REMARKS

The present invention provides methods for identifying active proteins in a complex protein mixture (*e.g.*, a proteomic mixture). Following reaction of the complex protein mixture with a single activity based probe (ABP), the resulting protein conjugates are proteolytically digested to provide probe-labeled peptides. While the skilled artisan would have expected, prior to the present invention, that proteolytic digestion would lead to a more complex protein mixture, in accordance with the present invention, it has been demonstrated that such proteolysis simplifies the complex protein mixture during subsequent analysis.

In preferred embodiments, ABPs are selected such that each active target protein forms a conjugate with a single ABP, preferably at a single discrete location in the target protein; thus, each conjugate gives rise to a peptide labeled with a single probe. Enrichment separation or identification of one or more ABP-labeled peptides may be achieved using liquid chromatography and/or electrophoresis. Mass spectrometry may be employed to identify one or more ABP-labeled peptides by molecular weight and/or amino acid sequence. In particularly preferred embodiments, sequence information derived from one or more of the ABP-labeled peptide(s) may be used to identify the protein from which the peptide was originally derived. Variations of these aspects can involve comparison of two or more proteomes, *e.g.*, with a single ABP, or, when analysis comprises mass spectrometry, probes having different isotopic compositions.

The invention provides enhanced simplicity and accuracy in identifying the active protein composition of a complex protein mixture. Using ABPs that bind to active target proteins, the analysis of complex protein mixtures is greatly simplified, particularly by providing ABPs that bind to active target proteins at a single site. The proteins are then proteolytically digested, resulting in a single representative ABP-labeled peptide fragment from each of the conjugates.

Using various approaches to identify the ABP-labeled peptide, the protein from which each ABP-labeled peptide originally was derived can be identified.

By the present communication, claims 21-24, 26-30 and 32 have been amended and new claims 49-74 have been added to define Applicant's invention with greater particularity. No new matter is introduced by the subject amendments as the amended claim language is fully supported by the specification and original claims. In addition, by the present communication, claims 1-20 and 33-47 have been cancelled without prejudice.

Upon entry of the amendments submitted herewith, claims 21-32 and 48-74 are currently pending. The present status of all claims in the application, and current amendments thereto, are provided in the Listing of Claims presented herein beginning on page 2.

This communication addresses all of the rejections raised in the most recent Office Action. Copies of the references cited in the prior Office Action response are attached hereto.

The rejection of claims 21, 22, 24, 25, 27, 28 and 48 under 35 U.S.C. § 102(e) as allegedly being anticipated by Cravatt et al., U.S. 2002/0045194, is respectfully traversed. Applicant's invention, as defined by amended claim 21, distinguishes over Cravatt by requiring a method for determining the presence, amount, or activity of one or more active target proteins in a complex protein mixture, the method consisting essentially of:

(a) contacting said complex protein mixture with a single activity based probe that specifically binds predominantly to a single target site on one or more active target proteins;

(b) optionally binding said target protein(s) to a solid support;

(c) proteolyzing said active target protein(s) to produce a product mixture;

(d) separating said product mixture into two or more components, one or more of which consist essentially of peptides bound to said probe; and

(e) generating a signal from said peptides bound to said probe, wherein said signal is correlated to the presence, amount, or activity of said one or more active target proteins in said complex protein mixture.

Thus, claim 21, as currently amended and presented for consideration, embraces only the steps or materials presented and those that do not materially affect the basic and novel characteristic(s) of the claim. For example, comparison of two or more complex protein samples can be done with a single activity based probe, consistent with the consisting essentially of language of claim 21. In contrast, for the comparison of two or more proteins, Cravatt requires the use of sets of probes (e.g., a light probe and a heavy probe; see paragraph [0128] at page 14 of Cravatt). Analysis according to the method provided by Cravatt also requires simultaneous analysis of samples treated with the sets of probes by mass spectrometry. Clearly, the ability to analyze and compare two or more protein samples by using a single probe distinguishes the present invention (as defined in claim 21) over Cravatt. The features of dependent claims 22-32 and 48 do not materially affect the steps of claim 21, depend from an allowable claim and thus should also be allowed.

Applicant's invention, as defined by newly added claim 49, further distinguishes over Cravatt by requiring a method for determining the presence, amount, or activity of one or more active target proteins in a complex protein mixture, the method comprising:

(a) contacting said complex protein mixture with a single activity based probe that specifically binds predominantly to a single target site on one or more active target proteins, wherein said probe comprises a fluorescent moiety;

(b) proteolyzing said active target protein(s) to produce a product mixture;

(c) separating said product mixture into two or more components, one or more of which comprise peptides bound to said probe, said probe using a receptor that specifically binds to said probe, wherein said receptor is an antibody or fragment thereof that binds to said fluorescent moiety; and

(d) generating a signal from said peptides bound to said probe, wherein said signal is correlated to the presence, amount, or activity of said one or more active target proteins in said complex protein mixture.

New claim 49 incorporates the features of previous claims 21-23. Since the prior Office Action acknowledged that claim 23 was not anticipated by Cravatt, new claim 49 is submitted to be allowable as currently presented. Specifically, new claim 49 includes the feature that the receptor is an antibody or fragment thereof that binds to the fluorescent moiety. This feature is not described by Cravatt. New claim 49 and claims 50–59 depending therefrom should therefore be allowed.

Applicant's invention, as defined in new claim 60, further distinguishes over Cravatt by requiring a method for comparing the presence, amount or activity of one or more active target proteins in each of two or more discrete proteomes, the method comprising:

(a) contacting each of said discrete proteomes with a single activity based probe that binds predominantly to a single target site on one or more active target proteins, wherein the same activity based probe is used for each discrete proteome;

(b) proteolyzing said discrete proteomes to produce a product mixture;

(c) separating each of said product mixtures into two or more components, one or more of which comprise peptides bound to said probe; and

(d) comparing the presence, amount or activity of the active target proteins in each of the discrete proteomes by generating a signal from said one or more components comprising peptides bound to said probe.

In contrast, when comparing multiple protein samples, Cravatt requires the use of sets of probes, isotopically related to each other in that one probe is heavier than the other. Cravatt does not disclose the use of a single probe for comparing protein samples. Thus, new claim 60 is allowable. Claims 61-73 depending therefrom should be allowed as depending from an allowable claim.

Applicant's invention, as defined by new claim 74, further distinguishes over Cravatt by requiring a method for detecting the presence, amount or activity of one or more active target proteins in a single complex protein mixture, the method comprising:

(a) contacting said complex protein mixture with an activity based probe that specifically binds predominantly to a single target site on one or more active target protein;

(b) proteolyzing said active target proteins to produce a product mixture;

(c) separating said product mixture into two or more components, one or more of which comprise peptides bound to said probe; and

(d) generating a signal from said peptides bound to said probe, wherein the signal is correlated to the presence, amount, or activity of said one or more active target proteins in said complex protein mixture.

Cravatt does not contemplate a method for analyzing proteins in a single complex protein mixture (e.g. a proteome) according to the above protocol. Protein digestion prior to separation is only described in Cravatt when comparing protein samples using sets of

isotopically labeled probes. See [0128]. In contrast, Cravatt teaches that in analyzing single proteins in complex mixtures, protein digestion is done after separation.

The rejection of claims 21-28, 30-32 and 48 under 35 U.S.C. § 103(a) as allegedly being unpatentable over Aebersold et al., U.S. 2002/0076739, in view of Cravatt, is respectfully traversed. As will be discussed below, the methods according to Aebersold teach away from the present invention. Because Aebersold does not teach an activity based probe that binds predominantly to one site, multiple peptides are produced creating a substantially more complex mixture than is initially provided. In contrast, the present invention provides a probe that binds substantially to one site, thereby leading to fewer peptides and a less complex separation process.

Applicant's invention, as defined, for example, by amended claim 21, distinguishes over the combination of Aebersold in view of Cravatt by requiring a method for determining the presence, amount, or activity of one or more active target proteins in a complex protein mixture, the method consisting essentially of:

(a) contacting said complex protein mixture with a single activity based probe that specifically binds predominantly to a single target site on one or more active target proteins;

(b) optionally binding said target protein(s) to a solid support;

(c) proteolyzing said active target protein(s) to produce a product mixture;

(d) separating said product mixture into two or more components, one or more of which consist essentially of peptides bound to said probe; and

(e) generating a signal from said peptides bound to said probe, wherein said signal is correlated to the presence, amount, or activity of said one or more active target proteins in said complex protein mixture.

Aebersold does not disclose or suggest such a method. Instead, as acknowledged by the Examiner, "Aebersold et al. differ from the instant invention in failing to teach the probe is an activity based probe." (See page 4 of the Office Action). This is a very significant difference. Activity based probes label a single target site on each protein, thus, following a proteolytic digest, only a single labeled peptide from each protein will be present. Prior to the present invention, the standard belief in the mass spectrometry community was that a single peptide did not provide data with sufficient confidence to unambiguously identify a protein through automated sequence searching algorithms.

With respect to the present invention as defined in claims 49-59, the remarks presented above with respect to claims 20-32 and 48 are reiterated. With respect to the present invention as defined in claims 60-74, it is respectfully submitted that the combined teachings of Abersold and Cravatt do teach or suggest the use of a single activity based probe in the comparison of two or more complex protein samples. As defined in the present invention, claim 60, step (a) requires "contacting each of said discrete proteomes with the same single activity based probe that binds predominantly to a single target site on one or more active target proteins." As previously stated with respect to the 102 rejection, comparison of complex protein samples according to Cravatt requires the use of sets of isotopically labeled probes. There is no teaching or suggestion to use a single activity based probe in comparing the samples. Thus, claims 60-74 are not obvious over Abersold in view of Cravatt.

The rejection of claim 29 under 35 U.S.C. § 103(a) as allegedly being unpatentable over Aebersold et al. and Cravatt, and further in view of Little et al., U.S. 2003/0003465, is respectfully traversed. Applicant's invention, as defined, for example, by claim 29, distinguishes over the combination of Aebersold and Cravatt in view of Little, by requiring a method for determining the presence, amount, or activity of one or more active target proteins in a complex protein mixture, the method consisting essentially of:

(a) contacting the complex protein mixture with a single activity based probe that specifically binds predominantly to a single target site on one or more active target proteins;

(b) optionally binding said target protein(s) to a solid support;

(c) proteolyzing the active target protein(s) to produce a product mixture, wherein prior to proteolyzing, the one or more active target proteins bound to the probe are bound to a solid support;

(d) separating the product mixture into two or more components, one or more of which consist essentially of peptides bound to the probe; and

(e) generating a signal from the peptides bound to the probe, wherein the signal is correlated to the presence, amount, or activity of the one or more active target proteins in the complex protein mixture.

As discussed above, neither Aebersold nor Cravatt, taken alone or in combination, are capable of rendering the present invention, as defined in the current claims, obvious. Indeed, as acknowledged by the Examiner, "Aebersold et al and Cravatt et al differ from the instant invention in failing to teach prior to the proteolyzing step, the [sp—that] one or more active target protein[(s)] bound to the probe are bound to a solid support." (See page 5 of the Office Action).

Further reliance on Little is unable to cure the deficiencies of the primary reference. No motivation to combine the cited references is provided in the Office Action. Indeed, it is respectfully submitted that the asserted combination of references can only be advanced with improper hindsight analysis, having benefit of Applicant's specification.

As discussed in substantial detail in Applicant's prior communication, when only a small number of peptides matched a particular protein (1-3 peptides), researchers manually inspected the MS data to determine test validity.[1] Since a typical mass spectrometry run generates on the order of 4000 spectra, manual analysis (at a rate of about 15 minutes per spectrum) would not be feasible if all data was expected to fall into the category of single site labeling of each protein.

In contrast to Aebersold, activity based probes label a single site on each target and proteolytic digestion does not increase the complexity of the APB labeled sample. The probes described by Aebersold teach away from the present invention, as the probes label multiple sites on each target protein.[2] Thus when a labeled sample according to Aebersold is digested, the number of labeled peptide species is substantially increased, typically at least 10X versus the number of labeled proteins. Because the methods of the present invention produce substantially fewer peptides, this allows for separation methods not applicable or possible with the Aebersold methods, e.g., lower resolution, higher throughput separation methods such as CE or LC (instead of LCMS/MS typically used for the Aebersold methods).

Furthermore, activity based probes of the present invention are typically larger than the Aebersold probes and the labeling sites of the probes result in very large peptides allowing for mass spectrometry data only on higher charge states of peptide ions (+3, +4, +5) to be collected. Previously, the belief was that only +1, +2, and in some cases +3 ions, provided data with

---

[1] See, for example, Florens et. al., *Nature* **419**, 520-26 (2002) (especially the final paragraph of the methods section, which discusses the analysis of "low coverage loci" by visual inspection), Adkins et. al., *Molecular and Cellular Proteomics*, December, 947-55 (2002)(specifically the legend of Table I legend (p. 949), which expressly states that "When three or fewer peptides for an individual protein passed the criteria shown in Table I, the mass spectra for those peptides were inspected manually."), and Washburn et. al., *Nature Biotechnology*, **19**, 242-47 (2001) (specifically the final paragraph of the experimental section, which states that "We manually confirmed each SEQUEST result from every protein identified by four or fewer peptides").

[2] Typically 5-30 sites—if, for example, the Aebersold "protein reactive group" is a cysteine, since the average protein in the complete human database has about 360 amino acids and cysteine has a relative abundance of 2.8%, the average number of cysteine residues per protein is 10.

sufficient quality to give a confident peptide sequence. Indeed, most laboratories only attempted to identify peptide sequences from +1 and +2 peptide data.[3]

Still further, specific labeling sites of activity based probes can be accurately predicted in most cases, thus the peptides potentially present in a digested, ABP-labeled sample can also be analyzed *in silico* (computationally). Even within families of closely related enzymes, peptides derived from tryptic digests of ABP labeled proteins have significant differences in their amino acid sequences enabling separation by standard chromatographic methods, and/or identification by mass spectrometry. Generally >95% of such peptides are non-redundant, i.e., the particular sequence is not shared by any other protein. This is neither taught or suggested by the cited prior art, and is a key realization for the success and/or general functionality of the claimed method.

Yet another prejudice in the art which Applicant had to overcome to arrive at the present invention was the fact that, at the time of the present invention, the ability to obtain consistent proteolytic digests in a multitude of proteomic mixtures had not been described or validated in the art. Nearly all published proteolytic digest procedures suggested using an amount of protease (trypsin) equal to a particular fraction of the amount of protein in the sample (e.g., 1 mg trypsin per 20 mg protein).[4]

---

[3] See, for example, Aebersold et al., *Chemical Reviews*, **101**, 269-95 (2001) (first paragraph on p. 278 states that "$[M+2H]^{2+}$ ions of peptides will produce tandem mass spectra of higher quality than those from either $[M+H]^{+}$ or $[M+3H]^{3+}$ peptide ions. The $[M+2H]^{2+}$ peptide ions fragmented under low-energy CID produce spectra...that are more readily interpreted than tandem mass spectra of $[M+3H]^{3+}$ and higher charge states"); Washburn et. al., *Nature Biotechnology*, **19**, 242-47 (2001), (the final paragraph of the experimental section states "Peptides identified by SEQUEST may have three different charge states (+1, +2, +3), each of which results in a unique spectrum for the same peptide."); Shen et. al., *Analytical Chemistry*, **76**, p 1134-44 (2004) (see the experimental section which provides the score criteria used for +1, +2, and +3 ions. Charge states higher than +3 were not searched); Florens et. al., *Nature*, **419**, 520-26 (2002) (see the experimental section thereof which provides the score criteria used for +1, +2, and +3 ions. Charge states higher than +3 were not searched); and Adkins et. al., *Molecular and Cellular Proteomics*, 947-55 (2002) (see the experimental section thereof which provides the score criteria used for +1, +2, and +3 ions. Charge states higher than +3 were not searched.)

[4] See, for example, Adkins et. al., *Molecular and Cellular Proteomics*, December, 947-55 (2002) (especially the experimental section at p 948 which states that samples were "digested with trypsin 1:50 (w/w) ratio

Researchers in the field at the time of the present invention believed that under such conditions trypsin was exhibiting first order kinetics (i.e., the rate of proteolysis was dependent only on the trypsin concentration, and the trypsin active site was saturated with substrate), which if true, would require a precise determination of protein concentration for every sample, limiting throughput and increasing the sampling requirements. Contrary to prior belief in the art and in accordance with the present invention, trypsin operates under second order kinetics at typical protein concentrations used in proteomics experiments (0.2-50 mg/mL). Thus when more protein is presented to the enzyme, turnover rate increases. Therefore, a constant amount of trypsin can be added to any sample, and the time required to reach a particular degree of proteolysis (e.g., 99% cleaved) is constant regardless of the protein concentration of the proteomic sample.

Reliance on the teaching of Cravatt is unable to cure the deficiencies of Aebersold. It is respectfully submitted that the combination of references does not teach the invention as presented in the amended claims, namely the analysis of a single proteome or the comparison of multiple proteins in a sample with a single activity based probe.

In view of the above amendments and remarks, reconsideration and favorable action on all claims are respectfully requested. In the event any issues remain to be resolved in view of this communication, the Examiner is invited to contact the undersigned by telephone so that a prompt disposition of this application can be achieved.

The Commissioner is hereby authorized to charge any additional fees which may be required regarding this application under 37 C.F.R. §§ 1.16-1.17, or credit any overpayment, to Deposit Account No. 50-0872. Should no proper payment be enclosed herewith, as by a check

---

for 2 h") and Shen et. al., *Analytical Chemistry*, **76**, 1134-44 (2004) (especially the experimental section, at p 1136, which states that the "protein was enzymatically digested using sequencing-grade modified porcine trypsin at a ratio of 1:50 (w/w)").

being in the wrong amount, unsigned, post-dated, otherwise improper or informal or even entirely missing, the Commissioner is authorized to charge the unpaid amount to Deposit Account No. 50-0872. If any extensions of time are needed for timely acceptance of papers submitted herewith, Applicant hereby petitions for such extension under 37 C.F.R. §1.136 and authorizes payment of any such extensions fees to Deposit Account No. 50-0872.

Respectfully submitted,

Date 4/26/05 By _____

FOLEY & LARDNER LLP
Customer Number: 30542
Telephone: (858) 847-6711
Facsimile: (858) 792-6773

Stephen E. Reiter
Attorney for Applicant
Registration No. 31,192

Enclosures

Adkins, et al., *Molecular & Cellular Proteomics*, December, 947-55 (2002).

Aebersold, et al., *Chem. Rev.*, 101, 269-95 (2001).

Florens, et al., *Nature*, 419, 520-26 (2002).

Shen, et al., *Anal. Chem.*, 76, 1134-44 (2004).

Washburn, et al., *Nature Biotechnology*, 19, 242-247 (2001).

Toward a Human Blood Serum Proteome

ANALYSIS BY MULTIDIMENSIONAL SEPARATION COUPLED WITH MASS SPECTROMETRY*[S]

Joshua N. Adkins‡, Susan M. Varnum‡, Kenneth J. Auberry§, Ronald J. Moore§, Nicolas H. Angell§¶, Richard D. Smith§, David L. Springer‡, and Joel G. Pounds‡‖

Blood serum is a complex body fluid that contains various proteins ranging in concentration over at least 9 orders of magnitude. Using a combination of mass spectrometry technologies with improvements in sample preparation, we have performed a proteomic analysis with submilliliter quantities of serum and increased the measurable concentration range for proteins in blood serum beyond previous reports. We have detected 490 proteins in serum by on-line reversed-phase microcapillary liquid chromatography coupled with ion trap mass spectrometry. To perform this analysis, immunoglobulins were removed from serum using protein A/G, and the remaining proteins were digested with trypsin. Resulting peptides were separated by strong cation exchange chromatography into distinct fractions prior to analysis. This separation resulted in a 3–5-fold increase in the number of proteins detected in an individual serum sample. With this increase in the number of proteins identified we have detected some lower abundance serum proteins (ng/ml range) including human growth hormone, interleukin-12, and prostate-specific antigen. We also used SEQUEST to compare different protein databases with and without filtering. This comparison is plotted to allow for a quick visual assessment of different databases as a subjective measure of analytical quality. With this study, we have performed the most extensive analysis of serum proteins to date and laid the foundation for future refinements in the identification of novel protein biomarkers of disease. ***Molecular & Cellular Proteomics 1:947–955, 2002.***

Serum, derived from plasma with clotting factors removed, contains 60–80 mg of protein/ml in addition to various small molecules including salts, lipids, amino acids, and sugars (1). The major protein constituents of serum include albumin, immunoglobulins, transferrin, haptoglobin, and lipoproteins (1, 2). In addition to these major constituents, serum also contains many other proteins that are synthesized and secreted, shed, or lost from cells and tissues throughout the body (3, 4). It is estimated that up to 10,000 proteins may be commonly present in serum, most of which would be present at very low relative abundances (5).

Historically, two-dimensional PAGE has been the primary method of separation and comparison for complex protein mixtures. This method has been critical in developing our understanding of the complexity and variety of proteins contained in cells and bodily fluids. Two-dimensional PAGE has been used to analyze serum and plasma (the unclotted parent fluid of serum) (6–13). Although impressive improvements in two-dimensional PAGE technologies have occurred in recent years, limitations remain. Two-dimensional PAGE is labor-intensive, requires relatively large sample quantities, is poorly reproducible, has a limited dynamic range for protein detection, and has difficulties in detecting proteins with extremes in molecular mass and isoelectric point (14). To address these limitations several types of mass spectrometry, in conjunction with various separation and analysis methods, are increasingly being adopted for proteomic measurements (15–22).

One of the driving forces in proteomics is the discovery of biomarkers, proteins that change in concentration or state in associations with a specific biological process or disease. Determination of concentration changes, relative or absolute, is fundamental to the discovery of valid biomarkers. The presence of higher abundance proteins (greater than mg/ml in serum) interferes with the identification and quantification of lower abundance proteins (lower than ng/ml in serum). Other methods such as two-dimensional PAGE have been used to demonstrate that the removal or separation of high abundance proteins enables greatly improved detection of lower abundance proteins (10, 11, 17, 23). The necessity of this removal or separation is also illustrated by noting that many proteins found useful as biomarkers for malignant and non-malignant disease (*e.g.* C-reactive protein, osteopontin, and prostate-specific antigen) are below 10 ng/ml, a value that is at least 7–8 orders of magnitude less than the most abundant serum proteins (1). Thus, the dynamic range typified by traditional proteomic methods are inadequate to allow for detection of these lower abundance serum proteins, or biomarkers, without effective removal or separation of the high abundance proteins.

One problem associated with any protein separation technique is that low abundance proteins may be removed along with the abundant species (24). Albumin is a protein of very high abundance in serum (35–50 mg/ml) that would be a prime candidate for complete selective removal prior to per-

From the ‡Biological Sciences Department and the §Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352

Received, October 2, 2002, and in revised form, November 13, 2002

Published, MCP Papers in Press, November 15, 2002, DOI 10.1074/mcp.M200066-MCP200

forming a proteomic analysis of lower abundance proteins. However, albumin is a transport protein in blood serum that binds a large variety of compounds including hormones, lipoproteins, and amino acids (1, 25, 26). Thus, removal of albumin from serum may also result in the specific removal of low abundance cytokines, peptide hormones, and lipoproteins of interest.

Immunoglobulins, or antibodies, are also abundant proteins in serum that function by recognizing "foreign" antigens in blood and initiating their destruction. To recognize this enormous variety of antigens present in blood, immunoglobulins contain variable regions (1, 25, 27). These variable regions are a source of random peptide sequence in serum that can complicate protein identifications from peptide sequences. Therefore, with immunoglobulins binding foreign materials and the random nature of sequences from their variable regions, removal of immunoglobulins is important for a proteomic analysis of serum.

The purpose of this investigation was to establish new preparative methods to remove or separate high abundance serum proteins and to apply new proteomic approaches that increase the dynamic range available for the identification and characterization of serum proteins. These methods include the use of protein A/G covalently bound to acrylamide beads to selectively remove immunoglobulins, described earlier as a significant source of sequence variability found in serum. Further, these methods include the separation of trypsin-digested peptides prior to mass spectrometric analysis using both strong cation exchange (SCX)[1] chromatography and capillary gradient reversed-phase liquid chromatography. This investigation identifies a large number of proteins (490) from a single (submilliliter) serum sample and further provides the foundation for future studies with clinically important disease states.

EXPERIMENTAL PROCEDURES

Human Blood Serum—The human blood serum was acquired from a healthy anonymous female donor (Donor No. M99869) (Golden West Biologicals, Temecula, CA). Immediately after collection, plasma was isolated from whole blood without anti-coagulants by centrifugation. The plasma supernatant was allowed to clot overnight at room temperature, and the clotted material was removed by centrifugation under sterile conditions. Upon receipt at our laboratory, the serum was aliquoted into 1-ml units and stored at −80 °C. In subsequent preparation steps, proteins were detected, and concentrations were estimated, where appropriate, using denaturing (SDS) polyacrylamide gel electrophoresis with GELCODE blue staining (Pierce catalog no. 24590), absorbance at 280 nm, and/or with a Bradford protein assay using bovine serum albumin (BSA) as a protein standard (24, 28).

Depletion of Serum Immunoglobulins and Trypsin Digestion—The immunoglobulins (Igs) were depleted by affinity adsorption chromatography using protein A/G. 500 μl of serum was diluted with an equal amount of 20 mM sodium phosphate, pH 8.0 and added to UltraLink Immobilized protein A/G beads (2:1, v/v) (Pierce) that had been equilibrated with 20 mM sodium phosphate, pH 8.0. This mixture was incubated with gentle rocking for 2 h at 25 °C. Immunoglobulin-depleted serum was separated from the protein A/G beads by centrifugation. The beads were washed three times with 5 volumes of PBS (150 mM NaCl, 10 mM sodium phosphate, pH 7.3), and the washes were pooled with the immunoglobulin-depleted serum. The diluted immunoglobulin-depleted serum sample was then dialyzed into 10 mM HCO_3NH_4, 5% acetonitrile, pH 7.5, digested with trypsin 1:50 (w/w) ratio (Promega, Madison, WI) for 2 h at 37 °C, and lyophilized.

Strong Cation Exchange Separation of Immunoglobulin-depleted Serum Peptides—Lyophilized, immunoglobulin-depleted serum peptides were resuspended in 2 ml of 75% 10 mM ammonium formate, 25% acetonitrile, pH 3.0 with formic acid. The sample was centrifuged to remove insoluble debris and then separated using an LC gradient ion exchange system consisting of a quaternary gradient pump (ThermoSeparations P4000, San Jose, CA) equipped with a polysulfoethyl A column (5 μm, 300 Å, PolyLC, Columbia, MD). Mobile phase A consisted of 75% 10 mM ammonium formate, 25% acetonitrile, pH 3.0 with formic acid, and mobile phase B was 75% 200 mM ammonium formate, 25% acetonitrile, pH 8.0. The column was initially loaded (2-ml injection loop) and equilibrated for 5 min with 0% B. Peptides were eluted using a linear gradient of 0–100% B over 30 min, and the column was subsequently washed at 100% B for an additional 25 min all at a flow rate of 4 ml/min. The column effluent was monitored at 280 nm with a Linear 200 UV detector (Micro-Tech Scientific, Sunnyvale, CA), and a total of 120 fractions were collected at 30-s intervals using a FRAC-100 (Amersham Biosciences). Collected fractions were lyophilized and stored at −80 °C for reversed-phase LC/MS/MS analysis.

Reversed-phase Separation and LCQ Ion Trap Analysis—Reversed-phase separation was performed with an Agilent 1100 capillary high pressure liquid chromatography system with a 60-cm capillary column (150-μm inner diameter × 360-μm outer diameter, Polymicro Technologies, Phoenix, AZ) packed with 5-μm Jupiter C_{18} particles (Phenomenex, Torrance, CA). Mobile phase A consisted of water and 0.1% formic acid, and mobile phase B consisted of acetonitrile and 0.1% formic acid. SCX fractions were dissolved in 50 μl of water, 0.1% formic acid. Peptides were injected on the column in 8 μl at a flow rate of 1.8 μl/min, and the column was re-equilibrated with 5% B for 20 min. Peptides were eluted with a linear gradient from 5 to 70% B over 80 min. The capillary column was interfaced to an LCQ Deca XP ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) using electrospray ionization.

The mass spectrometer was configured to optimize the duty cycle length with the quality of data acquired by alternating between a single full MS scan followed by three MS/MS scans on the three most intense precursor masses (as determined by Xcaliber mass spectrometer software in real time) from the single parent full scan. Dynamic mass exclusion windows were used and varied from 3 to 9 min. In addition, MS spectra for all samples were measured with an overall mass/charge (m/z) range of 400–2000. Fractions 21, 34, 39, 46, and 53, which contained high peptide concentrations, were re-analyzed three times using overlapping m/z ranges of 500–1050, 1000–1550, and 1500–2000, respectively. These segmented mass range analyses also utilized static mass exclusion lists that removed m/z precursors corresponding to the 20 most abundant peptides that were observed in the initial unsegmented analysis.

SEQUEST Analysis of Peptides—Tandem mass spectra were analyzed by SEQUEST (Bioworks 2.0, ThermoFinnigan) (16, 29–32), which performs its analyses by cross-correlating experimentally ac-

[1] The abbreviations used are: SCX, strong cation exchange; HUPO, Human Proteome Organization; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; NCBI, National Center for Biotechnology Information.

quired mass spectra with theoretical idealized mass spectra generated from a database of protein sequences. These idealized spectra are weighted largely with *b* and *y* fragment ions, *i.e.* fragments resulting from the amide linkage bond from the N and C termini, respectively. For these analyses, no enzyme rule restrictions were applied to the possible cleavage points available for peptide generation from the initial proteins, allowing identifications resulting from non-tryptic cleavage to be observed as well. The peptide mass tolerance was 3.0, and the fragment ion tolerance was 0.0.

Protein Databases—SEQUEST analysis was performed using a modified version of the human FASTA protein database provided with SEQUEST (ThermoFinnigan). Database modifications included the removal of viral proteins and the removal of some redundant protein entries as well as minimizing the number of entries for abundant serum proteins (13). Additional analyses were conducted using the National Center for Biotechnology Information (NCBI) human protein database[2] and the Unigene human database[3] to determine whether important abundant serum proteins were missing from our modified database. Use of the additional various human databases did not alter the vast majority of SEQUEST peptide identifications. The use of the larger databases did result in an expected decrease in magnitude of the SEQUEST DelCN score in a fraction of peptide identifications. Most peptides not found in the smaller supplied database did not pass subsequent filters including visual inspection of fragmentation spectra (data not shown), and in the case of the Unigene database analysis required up to 2 weeks to finish on a modern PC. Currently no complete human protein database has been compiled, and one is not likely to exist for a number of years (35). Thus, the modified database was considered to be an adequate resource for this initial blood serum proteome analysis after comparisons to the NCBI and Unigene databases.[2,3]

Of concern with a shotgun proteomic approach is whether assumptions made for simple cases continue to apply with higher levels of complexity. To address the question for database choice, we sought to analyze LC/MS/MS results using a smaller database containing very few peptides with sequence identity to human proteins but still retaining the level of complexity observed in a complete genome. A locally available *Deinococcus radiodurans* FASTA database derived from the open reading frames of a completely sequenced genome (15) was used to generate SEQUEST analyses to compare against the human database-derived results. Five SCX fractions (fractions 21, 34, 39, 46, and 53) that contained the greatest number of fully tryptic peptides were analyzed against the *D. radiodurans* database for this comparison.

Filters for SEQUEST Results—SEQUEST results were filtered (Table I) with criteria similar to those developed by Yates and co-workers (31, 36). Serum proteins in circulation are frequently found cleaved by chymotrypsin and elastase (37). Thus, while trypsin was used to digest the serum proteins, the SEQUEST data filter was modified to allow for identification of peptides resulting from both chymotrypsin and elastase cleavage sites. The chymotrypsin and elastase filter levels were derived by comparing the SEQUEST-identified tryptic peptides to the identified non-tryptic albumin peptides. The high abundance and globular nature of albumin represented a useful reference for defining non-tryptic filter parameters. The resulting filters were those that resulted in four or more hits for any non-tryptic albumin peptide. These filters further resulted in 33 non-tryptic cleavage sites of the 133 total albumin cleavage sites.

The final filter parameters used to determine cross-correlation (Xcorr) cut-off values took into account both the charge state of the peptide and the proteolytic cleavage rules as shown in Table I. Additionally, a minimum value of 0.1 was used for DelCN, indicating that SEQUEST was readily able to distinguish between its first and second choices for identification (32). When three or fewer peptides for an individual protein passed the criteria shown in Table I, the mass spectra for those peptides were inspected manually. Manual inspection was performed using four criteria generally accepted as means for assessment of spectral quality (16, 36). First, the spectrum quality must be acceptable with the peaks to be used in the determination clearly above the noise base line. Second, some continuity must be present among the *b* or *y* fragments, *i.e.* fragments for three or more adjacent amino acids. Third, if proline is predicted to be present, then the corresponding *y* fragment should give an intense peak. Last, unidentified intense peaks should be verified as being either doubly charged or simply the mass of the precursor with one or two of the terminal amino acids removed.

TABLE I
Conservative filter parameters for SEQUEST results

The spectra for proteins with three or fewer unique peptide hits that met these criteria were manually inspected before inclusion to the protein list. Each protein with three or fewer passing peptide identifications had an average of 33 identifications that did not pass the above criteria but scored better than a 1.5 Xcorr and had a DelCN of at least 0.05.

Charge	Xcorr	Peptide type
+1	≥1.9	Fully tryptic
+1	≥2.1	Chymotryptic and/or elastic
+1	≥2.2	Partially tryptic, chymotryptic, and/or elastic
+2	≥2.2	Fully tryptic
+2	≥2.4	Partially tryptic, chymotryptic, and/or elastic
+2	≥3.0	No protease rules
+3	≥3.75	Tryptic, chymotryptic, and/or elastic only

RESULTS

Protein A/G for Immunoglobulin Depletion—We found that protein A/G affinity adsorption chromatography depleted essentially all of the immunoglobulins from serum as assessed by SDS-polyacrylamide electrophoresis (Fig. 1). Analysis of serum by MS is complicated by the fact that abundant proteins impede measurement of less abundant proteins. In addition, the abundant serum immunoglobulins have regions of high sequence variability that may complicate an MS-based sequence analysis of serum-derived peptides. Thus, to increase the dynamic concentration range and confidence of determination it is critical to remove the immunoglobulins from the serum sample. The heavy and light chain portions of the immunoglobulins were removed when visualized with GelCode Blue Stain (Fig. 1, *Lane 3*). Albumin is also slightly depleted by the same procedure (Fig. 1, *Lane 4*). This depletion is unexpected in that during the production of the chimeric protein A/G the albumin binding site from protein G was removed (38).

Multidimensional Peptide Separation—Albumin and other abundant non-immunoglobulin proteins may also present problems for an MS analysis. Many published methods of albumin separation have resulted either in poor depletion or

[2] NCBI, Hs GenBank™ Protein Databases ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/.

[3] NCBI, Hs Unigene Contig Databases ftp.ncbi.nlm.nih.gov/repository/UniGene/.

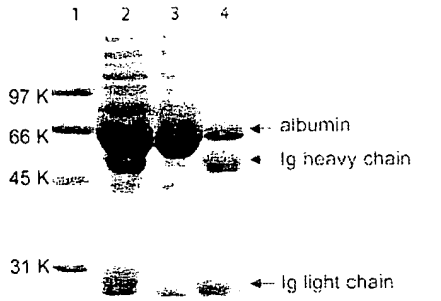

FIG. 1. **Igs, both heavy and light chain, are visibly depleted by protein A/G affinity chromatography as shown by SDS-PAGE.** Protein A/G was specific for immunoglobulins, but some cross-specificity for albumin was also present. *Lane 1*, molecular weight standards; *Lane 2*, unprocessed serum; *Lane 3*, serum after Ig depletion with protein A/G; *Lane 4*, proteins eluted from protein A/G.

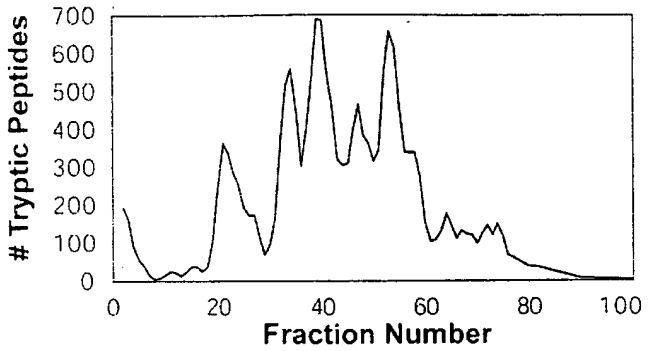

FIG. 2. **Strong cation exchange chromatography of human serum following depletion of immunoglobulins with protein A/G resulted in an excellent distribution of peptides over about 60 fractions (approximately fractions 19–79).** The number of fully tryptic peptides is as assessed by SEQUEST.

potential loss of specific low abundance proteins of interest in plasma (23) or in hemofiltrate (a plasma-derived fluid from dialysis patients) (17, 37). Rather than remove albumin from the serum, the strategy used here was to fractionate trypsin-derived peptides by SCX and then perform a second dimension separation with reversed-phase LC. The SCX chromatography resulted in good fractionation with the richest peptide samples eluted over about 60 fractions (fractions 19–79, Fig. 2).

The SCX fractionation illustrates the power of further analyzing specific fractions to increase the number of proteins determined by an LC/MS/MS analysis. Fractions 21, 34, 39, 46, and 53 were reanalyzed by LC/MS/MS using a static exclusion list for the 20 most commonly found peptides from the previous 400–2000 *m/z* MS analysis. In addition, each fraction was analyzed three times by limiting the *m/z* window to 500–1050, 1000–1550, or 1500–2000 for each run (illustrated in Fig. 3). The *m/z* segmentation resulted in approximately the same number of peptides passing SEQUEST data filters and manual inspection as the unsegmented analysis (Table II) but resulted in more proteins identified by multiple peptides and fewer numbers of serum albumin identifications. This increase in non-albumin identification is attributable to the MS analysis focusing on novel peptides rather than high abundance albumin peptides previously analyzed (Fig. 3). In addition, multidimensional separations allowed for important increases in dynamic range and decreases in individual analysis complexity. Here we show that some fractions may be complex enough to warrant further steps to simplify the MS analysis.

Proteins Identified in Serum—Using immunoglobulin depletion, SCX, and microcapillary reversed-phase high performance LC followed by data analysis with SEQUEST we have identified 490 proteins in serum. These proteins include those illustrated in Table III. Proteins found in this analysis also cover a large concentration range (as assessed from clinical reference normal values) from 85% coverage with 111 unique peptides from albumin (serum concentration 35–50 mg/ml), 31% coverage with 28 unique peptides from complement factor H (serum concentration 35 μg/ml); 29% coverage with 14 unique peptides from angiotensinogen (serum concentration 2.5–0.15 ng/ml) (39), 12% coverage with one peptide from prostate-specific antigen (serum concentration less than 1.0 pg/ml in a healthy female) (1). Our analysis identifies most serum proteins previously reported as well as a large number of proteins newly identified in serum (8–12, 37, 40).

Method for Visualizing and Accessing the Relative Quality of a Global SEQUEST Analysis—SEQUEST analysis results are typically scored using a combination of Xcorr and DelCN. Xcorr, in short, is the value of the best resulting correlation between a predicted peptide spectrum and an experimental spectrum. A higher Xcorr value provides better confidence of peptide identification. An Xcorr value greater than 2 is typically considered significant for peptide identifications. DelCN is the normalized difference in magnitude between the peptide fit with the highest Xcorr and the peptide fit with the second best Xcorr. A minimum acceptable value for DelCN is typically 0.1. More confidence is placed in protein identifications when multiple peptides occurring from the same protein that have Xcorr values greater than 2.0 and DelCN values greater than or equal to 0.1 (8, 16, 36).

To qualitatively evaluate the global results from a SEQUEST analysis, we compared the human peptides analyzed by MS/MS and *m/z* segmentation using SEQUEST with two different databases. The databases compared with SEQUEST analysis were an unrelated bacterial database (*D. radiodurans*) and a human protein database. The plot of DelCN *versus* Xcorr from a SEQUEST analysis with the *D. radiodurans* database generally defines a region of data that is composed of low confidence peptide identifications (Fig. 4*A*). A similar plot for a SEQUEST analysis using a human database identifies a second population of peptides with higher quality peptide identifications (Fig. 4*B*). The overlap between the poor quality and high quality populations contains many real peptide identifications. After filtering (see Table I), the SEQUEST analysis

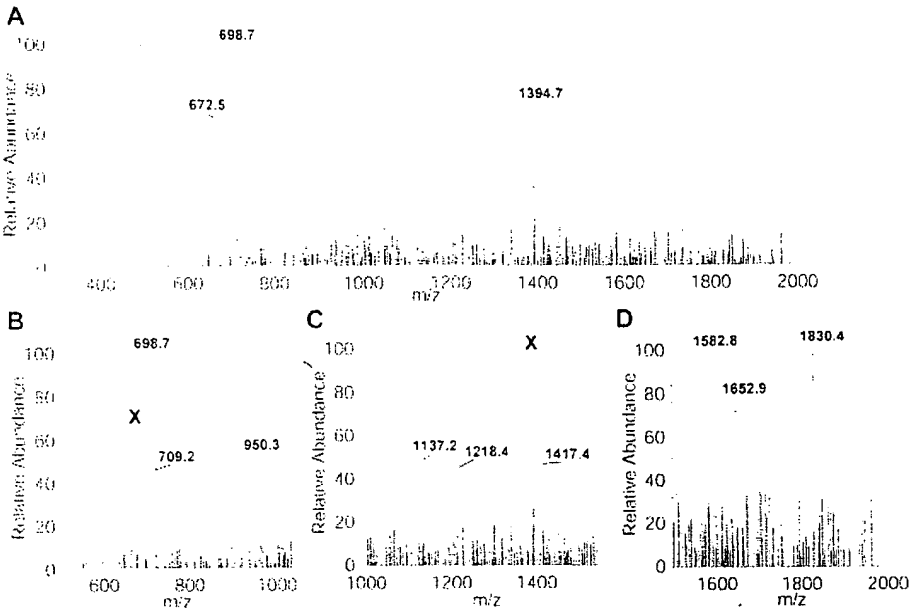

FIG. 3. **MS *m/z* segmentation illustration of the method used to measure peaks that may not have been missed during the initial analysis of SCX fractions 21, 34, 39, 46, and 53.** *A*, full scan with an *m/z* window of 300–2000; masses subsequently trapped and measured by tandem MS are labeled with *mass numbers*. *B*, *C*, and *D*, *m/z* segmented scans with *m/z* windows of 500–1050, 1000–1550, and 1500–2000, respectively. Masses analyzed by tandem MS are labeled with *mass numbers*, and masses disregarded due to static mass exclusion lists are labeled with an *X*.

TABLE II

Proteins found comparing a single MS/MS analysis to m/z window segmentation

Comparison of peptide information from a single analysis 400–2000 full MS scan of the five peak fractions (fractions 21, 34, 39, 46, and 53) (Fig. 2) and three *m/z* segmentations of the same five peak fractions with static exclusion list.

	400–2000 *m/z* full MS scan	500–1050, 1000–1550, and 1500–2000 *m/z* full MS scans
Total peptide hits	2014	2179
Proteins determined by ≥3 peptides[a]	26	58
Proteins determined by 1–2 peptides[a]	25	83

[a] Passing data filters in Table I.

of peptides using the *D. radiodurans* database eliminates all but 1% or 76 of the original low confidence peptide identifications (Fig. 4*C*). In contrast, after filtering (Table I), 20% or 2179 of the peptides from the human database remain (Fig. 4*D*). The filtering method results in more qualitative confidence for the peptide identifications using the human protein database at a global scale. While it is expected that most of the peptides identified from the *D. radiodurans* database that passed the data filters do not appear as proteins serum, some of these peptides may, by chance or evolutionary conservation, be legitimately found using the *D. radiodurans* database.

DISCUSSION

Blood plasma, like cells, has many high abundance proteins that perform various housekeeping functions. Blood plasma contains numerous secreted or shed low abundance proteins that are critical for signaling cascades and regulatory events. During necrosis, apoptosis, and hemolysis, contents of cells may be released into the plasma. The presence of these components in blood reinforces the benefits of using a proteomic approach for identifying biomarkers for disease states. In this study, we report an analysis of serum identifying 490 proteins (Table III and supplemental data table), at least a 3–5-fold increase in the number of identified proteins from a blood-derived fluid found in previous reports.

Previous proteomic characterizations of human plasma have used two-dimensional PAGE. These studies such as the seminal work of Anderson and co-workers (10, 41) have been summarized by the ExPASy on-line human plasma two-dimensional PAGE database (ca.expasy.org/ch2d/). These previous investigations have focused on plasma and thus are not directly comparable to the serum results reported here. However, of the 58 named proteins identified in this on-line human plasma protein database, we identified 51 in our serum analysis. There are several possible explanations for not identifying these seven proteins, including fibrinogen B, fibrinogen γ, C-reactive protein, and actin. First, plasma but not serum samples contain the clotting factors fibrinogen B and fibrinogen γ. Second, our serum was obtained from a single healthy female. The concentration of certain blood proteins may make detection difficult for our single source sample *versus* a general population; an example is C-reactive protein, which is typically at subnanogram per milliliter concentrations in a healthy female (1). Finally, the sample preparation and analytical methods used by these previous investigators differ significantly from those reported here. The lack of detection for the other proteins, such as actin, may be due to differing methods of sample collection, processing, and analysis. Overall our approach is superior for global identification since the two-dimensional PAGE database is made up of nine published reports but identified only 58 proteins, while we found

TABLE III

Selected 134 categorized proteins from the 490 total proteins detected

MAP, mitogen-activated protein; ERK, extracellular signal-regulated kinase.

Common circulating blood proteins	Albumin, haptoglobin, hemopexin, fibrinogen A, α_1-microglobulin, β_2-microglobulin, α_2-glycoprotein(Zn), α_2-HS-glycoprotein, serum amyloid proteins (A2-β and A), vitronectin, apolipoproteins (A-I, A-II, A-IV, B, C-I, C-II, C-III, D, E, F, and L), gelsolin, histidine-rich glycoprotein, leucine-rich α_2-glycoprotein, low density lipoprotein-related proteins (1 and 2), α_1-acid glycoprotein 1 (orsomucoid 1), α_1-acid glycoprotein 2 (orsomucoid 2), clusterin, Kell blood group protein, perlecan (heparan sulfate proteoglycan), ferroxidase
Coagulation and complement factors	Complement factors (B, C1R, C2, C3, C4A, C4B, C5, C6, C7, C8α, C8β, C8γ, C9, H, and I), coagulation factors (II, V, VIII, XII, and XIIIb)
Blood transport and binding proteins	Transferrin, transthyretin, retinol-binding protein, vitamin D-binding protein, insulin-like growth factor-binding proteins (5 and 7), calcium-binding protein P22, complement C4-binding protein α, hemoglobins (A and B), high density lipoprotein-binding protein, histidine-rich calcium-binding protein, hyaluronan-binding protein 2, latent transforming growth factor-β-binding protein, S100 calcium-binding protein A2, thyroglobulin, corticosteroid-binding globulin, selenoprotein P
Protease inhibitors	α_2-Antiplasmin inhibitor, complement C1 inhibitor, heparin cofactor II (protease inhibitor leuserpin 2), inter-α-(globulin) inhibitor H4 (plasma kallikrein-sensitive glycoprotein), inter-α-trypsin inhibitor, plasminogen activator inhibitor, protease inhibitor 4 (kallistatin), α_1-antichymotrypsin, α_1-antitrypsin
Proteases	Kallikrein, angiotensinogen, plasminogen, α-thrombin, carboxypeptidase N
Other enzymes	Antioxidant protein 1, arginosuccinase, hexokinase 3, folate hydrolase 1 (prostate-specific membrane antigen), nicotinamide nucleotide transhydrogenase, paraoxonase/arylesterase, phosphodiesterase 5A, phosphoglycerate kinase 1, squalene monooxygenase, triacylglycerol lipase, methylmalonyl coenzyme A mutase, thioredoxin-dependent peroxide reductase
Cytokines and hormones	Atrial natriuretic factor, human growth hormone, inhibin, interleukin-12a, interferon (α-inducible protein 27), fibroblast growth factor-12, prostate-specific antigen, growth/differentiation factor 5, pigment epithelium-derived factor
Channel and receptor-derived peptides	ATP-sensitive inward rectifier K^+ channel 11, chemokine (CX_3C) receptor 1, G protein-coupled receptor 1, γ-aminobutyric acid receptor B, prostaglandin E receptor (subtype EP3), protein tyrosine phosphatase (receptor type, f polypeptide), solute carrier family 5 (sodium iodide symporter), T-cell receptor α chain VJ region, tumor necrosis factor receptor-associated factor 5, interleukin-2 receptor γ chain, integrin α (4, 8, and E)
Miscellaneous (structural, nuclear, etc.)	Keratins (1, 2, and 9), microtubule-associated protein, microtubule-vesicle linker clip-170, plectin, sytaxin, elastin, MAP/ERK kinase kinase 5, bullous pemphigoid antigen, centromere protein f, collagens (IV and XI), titans, elongation factor tu, epidermal growth factor receptor pathway substrate 8

the 490 proteins, including those that would be expected to be common between studies.

Another family of serum/plasma studies for comparison is the characterization of rat serum by Gianazza and co-workers (6–8, 11, 12). These studies identified 34 proteins with human homologues and characterized the changes in protein abundance with disease states or chemical exposures associated with inflammatory disease. These rat serum studies concluded that even abundant proteins could be markers for disease states. Our study identified the human homologues of 31 of the 34 identified rat proteins. We did not find the human equivalent of thyroxine-binding globulin, thiostatin, or C-reactive protein. Many of the same reasons for a lack of complete overlap with the ExPASy plasma two-dimensional PAGE database apply here. In addition, species-specific differences may explain differing proteins and expression levels.

Serum is a complex biological fluid with many functions, and the presence, absence, or concentration of a specific protein may be non-intuitive until the serum proteome is fully understood. In an analysis of this complexity, it is important to note that expectations often differ from results for many proteins. Examples of unexpected results are hemoglobin and actin, which are both ubiquitous in the red blood cells. Therefore between high quantity and rapid turnover of red blood cells it may be expected that hemoglobin and actin should be readily detectable in serum (42). In contrast to our expectations, few hemoglobin-derived peptides and no actin-derived peptides were identified. In fact, both hemoglobin and actin are actively sequestered and cleared from the serum via the abundant serum proteins haptoglobin and vitamin D-binding protein, respectively (42–45). Another example of unexpected results are the identification of immunoglobulin-derived peptides, although depletion was complete when evaluated by SDS-PAGE. It is unclear whether these peptides originated from incomplete depletion of immunoglobulins *in vitro* or from proteolyzed immunoglobulins circulating in blood.

As global proteomic approaches become more common, there is an increasing need to evaluate and visualize large data sets with improvements in individual scoring methods (46–48). Often proteomic studies are less concerned with individual peptide identifications than with globally studying changes. In fact, a recent study using a global approach to profile proteins only by masses using surface-enhanced laser desorption-ionization MS with blood serum has been shown

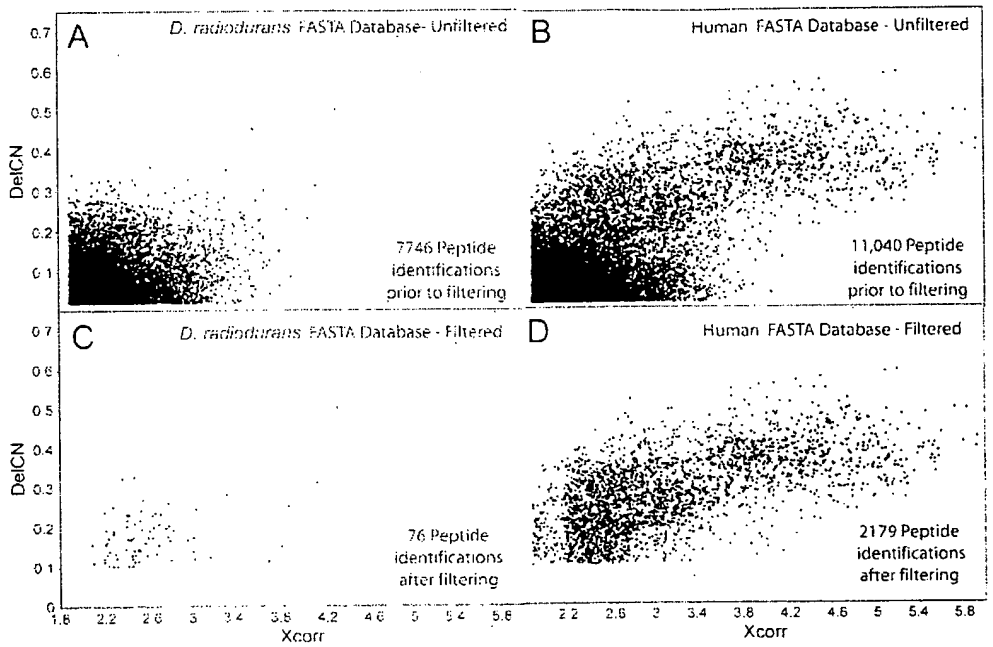

FIG. 4. **Global affects of SEQUEST peptide identification filters.** Shown are populations using a "random sequence" database with genome level complexity and a human protein database. The *D. radiodurans* database was used here as the random sequence database with results unfiltered (*A*) and filtered (*C*). A database derived from NCBI human protein sequence data was used as the human protein database with results unfiltered (*B*) and filtered (*D*).

to have predictive value in ovarian cancer (33). One of the difficulties related to the use of SEQUEST for peptide identifications is the lack of methods to globally evaluate the quality of data and the lack of methods to access global changes created by filtering schemes and/or database changes. Here, by comparing our SEQUEST results to multiple databases, we have illustrated an intuitive and easily adopted method for analyzing LC-MS/MS experiments in global terms (Fig. 4).

Major technical issues complicate the routine characterization of the plasma/serum proteome. First, plasma/serum proteins, like tissue proteins, may be post-translationally modified, and many plasma proteins are glycosylated (13). Other important factors include modifications such as sulfation, phosphorylation, oxidation, glycation, lipidation, and γ-carboxyglutamylation. Currently there are no commercially available tools that can identify peptides with this variety and number of modifications. The serum proteins in this study (Table III) were identified from translationally unmodified peptides. Significant improvements to sample processing and informatics are needed to identify these protein modifications. Second, protease digestion further adds to the complexity of a proteomic analysis of serum (13). Here we filtered peptide identifications based on protease modifications to take *in situ* proteolysis (chymotrypsin and elastase) into account. Third, the concentration range of plasma/serum proteins encompasses at least 9 orders of magnitude. Thus, significant improvements in the sample processing and separation with improvement in the dynamic range, sensitivity, and ability to quantitate results from mass spectrometry are needed to elaborate the plasma/serum proteome beyond the 490 proteins identified in this report. Last, the immature status of human protein databases further complicates analysis because there are likely to be protein identifications even in this mid-abundance range that have not yet been added to any publicly available human protein database (35).

The Human Proteome Organization (HUPO) has been founded to consolidate and organize future efforts in human proteomics (34). Among the many of the stated goals of HUPO are the research goals of characterizing the human plasma/serum proteomes and the informatic goals of standardizing proteome data and annotations with the improvement of bioinformatic tools for proteome analysis (34). Here we report a large improvement for proteomic analysis of serum; this analysis identifies 490 proteins, about 10% toward a 5000 protein goal of HUPO. Further, we have presented a visualization method that can be used to evaluate the quality of a global SEQUEST proteomic analysis along with the ability to subjectively evaluate protein database quality for a SEQUEST analysis.

Acknowledgments—We gratefully acknowledge the insightful discussions of David Wunschel and Richard Zangar, the technical assistance of Deanna Auberry, and the encouragement and support of Robert Miller and David Koppenaal.

* This work was supported by the Biotechnology section of Core Technology, Battelle Memorial Institute. The costs of publication of this article were defrayed in part by the payment of page charges.

This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains supplemental data: peptide identifications.

¶ Current address: Human Genome Sciences, 9410 Key West Ave., Rockville, MD 20850.

‖ To whom correspondence should be addressed: Biological Sciences Dept., Pacific Northwest National Laboratory, P.O. Box 999, MSIN: P7-58, Richland, WA 99352. Tel.: 509-376-1015; Fax: 509-376-9449; E-mail: joel.pounds@pnl.gov.

REFERENCES

1. Burtis, C. A., and Ashwood, E. R. (2001) *Tietz Fundamentals of Clinical Chemistry*, 5th Ed., W. B. Saunders Company, Philadelphia, PA
2. Turner, M. W., and Hulme, B. (1970) *The Plasma Proteins: An Introduction*, Pitman Medical & Scientific Publishing Co., Ltd., London
3. Schrader, M., and Schulz-Knappe, P. (2001) Peptidomics technologies for human body fluids. *Trends Biotechnol.* **19,** S55–S60
4. Kennedy, S. (2001) Proteomic profiling from human samples: the body fluid alternative. *Toxicol. Lett.* **120,** 379–384
5. Wrotnowski, C. (1998) The future of plasma proteins. *Genet. Eng. News* **18,** 14
6. Eberini, I., Agnello, D., Miller, I., Villa, P., Fratelli, M., Ghezzi, P., Gemeiner, M., Chan, J., Aebersold, R., and Gianazza, E. (2000) Proteins of rat serum V: adjuvant arthritis and its modulation by nonsteroidal anti-inflammatory drugs. *Electrophoresis* **21,** 2170–2179
7. Eberini, I., Miller, I., Zancan, V., Bolego, C., Puglisi, L., Gemeiner, M., and Gianazza, E. (1999) Proteins of rat serum IV. Time-course of acute-phase protein expression and its modulation by indomethacine. *Electrophoresis* **20,** 846–853
8. Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, M. R., Manzoni, C., Vignati, M., and Gianazza, E. (1998) Proteins of rat serum: I. establishing a reference two-dimensional electrophoresis map by immunodetection and microbore high performance liquid chromatography-electrospray mass spectrometry. *Electrophoresis* **19,** 1484–1492
9. Edwards, J. J., Anderson, N. G., Nance, S. L., and Anderson, N. L. (1979) Red cell proteins. I. two-dimensional mapping of human erythrocyte lysate proteins. *Blood* **53,** 1121–1132
10. Anderson, L., and Anderson, N. G. (1977) High resolution two-dimensional electrophoresis of human plasma proteins. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5421–5425
11. Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., and Gianazza, E. (1998) Proteins of rat serum: II. influence of some biological parameters of the two-dimensional electrophoresis pattern. *Electrophoresis* **19,** 1493–1500
12. Miller, I., Haynes, P., Eberini, I., Gemeiner, M., Aebersold, R., and Gianazza, E. (1999) Proteins of rat serum: III. gender-related differences in protein concentration under baseline conditions and upon experimental inflammation as evaluated by two-dimensional electrophoresis. *Electrophoresis* **20,** 836–845
13. Peters, T., Jr. (1987) Intracellular precursor forms of plasma proteins: their functions and possible occurrence in plasma. *Clin. Chem.* **33,** 1317–1325
14. Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2,** 3–10
15. Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Pasa-Tolic, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Anal. Chem.* **73,** 2132–2139
16. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
17. Raida, M., Schulz-Knappe, P., Heine, G., and Forssmann, W. G. (1999) Liquid chromatography and electrospray mass spectrometric mapping of peptides from human plasma filtrate. *J. Am. Soc. Mass Spectrom.* **10,** 45–54
18. Liotta, L. A., Kohn, E. C., and Petricoin, E. F. (2001) Clinical proteomics: personalized molecular medicine. *J. Am. Med. Assoc.* **286,** 2211–2214
19. Smith, R. D. (2000) Evolution of ESI-mass spectrometry and Fourier transform ion cyclotron resonances for proteomics and other biological applications. *Int. J. Mass Spectrom.* **200,** 509–544
20. Yates, J. R., III (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16,** 5–8
21. Wu, S.-L., Amato, H., Biringer, R., Choudhary, G., Shieh, P., and Hancock, W. S. (2002) Targeted proteomics of low-level proteins in human plasma by LC/MSn: using human growth hormone as a model system. *J. Proteome Res.* **1,** 459–465
22. Bergquist, J., Palmblad, M., Wetterhall, M., Hakansson, P., and Markides, K. E. (2002) Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrom. Rev.* **21,** 2–15
23. Georgiou, H. M., Rice, G. E., and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1,** 1503–1506
24. Scopes, R. K. (1994) *Protein Purification: Principles and Practice*, 3rd Ed., Springer-Verlag, New York
25. Ritchie, R. F., and Navolotskaia, O. (eds) (1996) *Serum Proteins in Clinical Medicine*, 1st Ed., Vol. 1, Foundation for Blood Research, Scarborough, ME
26. Beutler, E., and Williams, W. J. (1995) *Williams Hematology*, 5th Ed., McGraw-Hill Inc. Health Professions Division, New York
27. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
28. Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254
29. Yates, J. R., III, Carmack, E., Hays, L., Link, A. J., and Eng, J. K. (1999) Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry. *Methods Mol. Biol.* **112,** 553–569
30. Yates, J. R., III, McCormack, A. L., and Eng, J. K. (1996) Mining genomes with MS. *Anal. Chem.* **68,** 534–540
31. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,**242–247
32. Eng, J. K., Mccormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
33. Petricoin, E. F., Ardenkani, A. A., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
34. Hanash, S., and Celis, J. (2002) The human proteome organization: a mission to advance proteome knowledge. *Mol. Cell. Proteomics* **1,** 413–414
35. Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30,** 1083–1090
36. Wolters, D. A., Washburn, M. P., and Yates, J. R., III (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690
37. Richter, R., Schulz-Knappe, P., Schrader, M., Standker, L., Jurgens, M., Tammen, H., and Forssmann, W. G. (1999) Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J. Chromatogr. B Biomed. Sci. Appl.* **726,** 25–35
38. Pierce Endogen (1995) *Vol. 0497 Instructions: UltraLink Immobilzed Protein A/G*, pp. 1–4, Pierce Endogen, Rockford, IL
39. Vinck, W. J., Fagard, R. H., Vlietinck, R., and Lijnen, P. (2002) Heritability of plasma renin activity and plasma concentration of angiotensinogen and angiotensin-converting enzyme. *J. Hum. Hypertens.* **16,** 417–422
40. Eckerskorn, C., Strupat, K., Schleuder, D., Hochstrasser, D., Sanchez, J. C., Lottspeich, F., and Hillenkamp, F. (1997) Analysis of proteins by direct-scanning infrared-MALDI mass spectrometry after 2D-PAGE separation and electroblotting. *Anal. Chem.* **69,** 2888–2892
41. Hoogland, C., Sanchez, J. C., Tonella, L., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. (1999) The SWISS-2DPAGE database: what has changed during the last year. *Nucleic Acids Res.* **27,** 289–291
42. Houmeida, A., Hanin, V., Constans, J., Benyamin, Y., and Roustan, C. (1992) Localization of a vitamin-D-binding protein interaction site in the COOH-terminal sequence of actin. *Eur. J. Biochem.* **203,** 499–503
43. Emerson, D. L., Galbraith, R. M., and Arnaud, P. (1984) Electrophoretic

demonstration of interactions between Gc (vitamin D-binding protein), actin and 25-hydroxycholecalciferol. *Electrophoresis* **5**, 22–26

44. Goldschmidt-Clermont, P. J., Van Baelen, H., Bouillon, R., Shook, T. E., Williams, M. H., Nel, A. E., and Galbraith, R. M. (1988) Role of group-specific component (vitamin D binding protein) in clearance of actin from the circulation in the rabbit. *J. Clin. Investig.* **81**, 1519–1527
45. Haddad, J. G., Hu, Y. Z., Kowalski, M. A., Laramore, C., Ray, K., Robzyk, P., and Cooke, N. E. (1992) Identification of the sterol- and actin-binding domains of plasma vitamin D binding protein (Gc-globulin). *Biochemistry* **31**, 7174–7181
46. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6**, 207–212
47. MacCross, M. J., Wu, C. C., and Yates, J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74**, 5593–5599
48. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47

Mass Spectrometry in Proteomics

Ruedi Aebersold* and David R. Goodlett

The Institute for Systems Biology, 4225 Roosevelt Way NE, Seattle, Washington 98105

Received July 24, 2000

Contents

I. Introduction

Proteomics can be viewed as an experimental approach to explain the information contained in genomic sequences in terms of the structure, function, and control of biological processes and pathways. Proteomics attempts to study biological processes comprehensively by the systematic analysis of the proteins expressed in a cell or tissue. Mass spectrometry (MS) is currently proteomic's most important tool.

A. Genomics and Proteomics

The classical biochemical approach to study biological processes has been based on the purification to homogeneity by sequential fractionation/assay cycles of the specific activities that constitute the process; the detailed structural, functional, and regulatory analysis of each isolated component; and the reconstitution of the process from the isolated components. The Human Genome Project and other genome sequencing programs are turning out in rapid succession the complete genome sequences of specific species and thus, in principle, the amino acid sequence of every protein potentially encoded by that species.[1,2] It is to be expected that this information resource (unprecedented in the history of biology) will enhance traditional research methods such as the biochemical approach and also catalyze fundamentally different research paradigms, one of which is proteomics.[3–5]

The programs to sequence the entire human genome along with the genomes of a number of other species have been extraordinarily successful. The genomes of 46 microbial species (TIGR Microbial Database; www.tigr.org) have been completed, and the genomes of over 120 other microbial species are in the process of being sequenced. Additionally, the more complex genomes of eukaryotes, in particular those of the genetically well-characterized unicellular organism *Saccharomyces cerevisiae* and the multicellular species *Caenorhabditis elegans* and *Drosophila melanogaster* have been sequenced completely; a "draft sequence" of the rice genome has been published; and completion of the human and arabidopsis (92% complete in May 2000) genomes appear imminent.[6–11] Even in the absence of complete genomic sequences, rich DNA sequence databases have been publicly available, including those containing over 2.1 million human and over 1.2 million murine expressed sequence tags (ESTs).[12] ESTs are stretches of approximately 300−500 contiguous nucleotides representing partial gene sequences that are being generated by systematic single pass sequencing of the clones in cDNA libraries. On the time scale of most biological processes, with the notable exception of evolution, the genomic DNA sequence can be viewed as static. A genomic sequence database therefore represents an information resource akin to a library. Intensive efforts are underway to assign "function" to individual sequences in sequence databases. This is attempted by the analysis of linear sequence motifs or higher order structural motifs that indicate a statistically significant similarity of a sequence to a family of sequences with known function or by other means such as comparison of homologous protein functions across species.[13–17] These efforts will lead

* Corresponding author telephone: (206)732-1200; fax: (206)732-1299; e-mail: ruedi@systemsbiology.org.

10.1021/cr990076h CCC: $36.00
Published on Web 01/23/2001

Dr. Aebersold is a founding member of the Institute for Systems Biology in Seattle, WA, where he leads a research effort that is focused on developing new methods and technologies for understanding the structure, function, and control of complex biological systems. He completed his undergraduate studies in biology at the University of Basel, Switzerland, in 1979 and received a Ph.D. in Cell Biology at the University of Basel in 1984. Holding fellowships from the Swiss National Science Foundation and EMBO, he joined the California Institute of Technology as a postdoctoral fellow (1984–1986) and remained at Caltech as a senior research fellow (1986–1988). In 1988, he joined the University of British Columbia in Vancouver as an assistant professor in the Department of Biochemistry and Molecular Biology and as a senior investigator at the Biomedical Research Centre. In 1993, he moved to the University of Washington as an associate professor in Molecular Biotechnology and was promoted to full professor in 1998. In 2000, he left the University of Washington and joined the Institute for Systems Biology as co-founder and full faculty member. His research and teaching have been recognized by a long-term fellowship from the European Molecular Biology Organization (EMBO), by a scholarship from the Swiss National Science Foundation, by the Killam Research Prize, and by the Pehr Edman Award. He is a senior editor for the journal *Physiological Genomics*, has been a member of the Editorial Advisory Boards of *Protein Science* (1992–1998), *Functional Proteomics* (1999–present), *Analytical Biochemistry* (1991–present), *Functional and Integrative Genomics* (1999–present), and *Electrophoresis* (1989–1993).

Dr. Goodlett is a Senior Research Scientist and Director of the Proteomics Laboratory at the Institute for Systems Biology in Seattle, WA. Prior to this, he took an extended sabbatical (1998–1999) from the pharmaceutical industry to work with Prof. Ruedi Aebersold in the Department of Molecular Biotechnology at the University of Washington. During this time, he developed sensitive analytical methods for detection and sequencing of phosphopeptides. While employed in the pharmaceutical industry (1993–1997), he carried out analytical research on an HIV therapeutic for Johnson & Johnson, Inc. and drug discovery research in the field of immunology for Bristol-Myers Squibb, Inc. He is an expert in the field of protein/peptide characterization by mass spectrometry and in microscale separation sciences. Dr. Goodlett was a NORCUS Postdoctoral Fellow in the laboratory of Richard D. Smith at Battelle-Memorial Institute (1991–1992) where he developed mass spectrometric methods for determination of the thermodynamic properties of protein–protein complexes and methods to increase the sensitivity of capillary electrophoresis-mass spectrometry. He obtained a Ph.D. in Biochemistry from North Carolina State University (1991) in the protein mass spectrometry laboratory of Richard B. van Breemen and holds M.S. (1988) and B.S. (1982) degrees in Chemistry from Auburn University.

to a more richly annotated sequence database and, not by themselves, to an explanation of the structure, function, and control of biological processes.

The proteome has been defined as the protein complement expressed by a genome.[18–21] This somewhat restrictive definition implies a static nature of the proteome. In reality, the proteome is highly dynamic; the types of expressed proteins, their abundance, state of modification, subcellular location, etc. being dependent on the physiological state of the cell or tissue. Therefore, the proteome reflects the cellular state or the external conditions encountered by a cell, and proteome analysis can be viewed as a genome-wide assay to differentiate and study cellular states and to determine the molecular mechanisms that control them.[22] Considering that the proteome of a differentiated cell is estimated to consist of thousands to a few ten-thousands of different types of proteins with an estimated dynamic range of expression of at least 5 orders of magnitude, the prospects for proteome analysis appear daunting. However, the availability of (genomic) DNA databases listing the sequence of every potentially expressed protein and rapid advances in technologies capable of identifying the proteins that are actually expressed now make proteomics a realistic proposition. MS is one of the essential legs on which current proteomics technology stands.

B. MS and Proteomics

During the decade of the 1990s, changes in MS instrumentation and techniques revolutionized protein chemistry and fundamentally changed the analysis of proteins. These changes were catalyzed by two technical breakthroughs in the late 1980s: the development of the two ionization methods electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI).[23–25] These methods solved the difficult problem of generating ions from large, nonvolatile analytes such as proteins and peptides without significant analyte fragmentation. Because of the lack or minimal extent of analyte fragmentation during the ESI and MALDI processes, they are also referred to as "soft" ionization methods. In fact they are so soft that under specific conditions even noncovalent interactions may be maintained during the ionization process. ESI gained immediate popularity because of the ease with which it could be interfaced with popular chromatographic and electrophoretic liquid-phase separation techniques and quickly supplanted fast atom bombardment as the ionization method of choice for protein and peptide samples dissolved in a liquid phase.[26] Furthermore, due to the propensity of ESI to produce multiply charged analytes, simple quadrupole instruments and other types of mass analyzers with limited *m/z* range could be used to detect analytes with masses exceeding the nominal *m/z* range of the instrument. For different but no less compelling reasons, MALDI also rapidly gained popularity. The

time-of-flight (TOF) mass analyzer most commonly used with MALDI is robust, simple, and sensitive and has a large mass range. MALDI mass spectra are simple to interpret due to the propensity of the method to generate predominantly singly charged ions. The method is relatively resistant to interference with matrixes commonly used in protein chemistry.

Direct measurement of the molecular weight of large [>10 000 mass units (u)] polypeptides was quickly demonstrated with both ESI and MALDI.[25,27] More recently, the mass determination of very large proteins in excess of 100 000 u has been realized.[28,29] While the early roles of ESI-MS and MALDI-MS in protein analysis were essentially those of accurate balances,[30] the ease with which proteins and peptides could be ionized by these methods rapidly made MS a complementary technique to nuclear magnetic resonance, X-ray crystallography, circular dichroism, and the classical methods of protein chemistry to study diverse aspects of protein structure and function. Numerous reports document the success MS has enjoyed in studies into the four structural classifications of proteins, namely, the *primary* structure or linear sequence of amino acids, the *secondary* structure or the folding of stretches of amino acids into defined structural motifs, the *tertiary* structure or the overall three-dimensional fold, and the *quaternary* structure or the spatial arrangement of folded polypeptides in multiprotein complexes. The application of MS to proteomics, the subject of this review, has to date been realized mostly for the study of protein primary structures. However, the following anecdotal examples hint at an increasing role of MS in the systematic study of protein higher order structures, i.e., structural proteomics, as well as of protein–ligand interactions.

Because of their relative softness of ionization, ESI and MALDI have been used in attempts to generate gas-phase ions of noncovalently associated, apparently intact protein complexes for the purpose of studying these structures by MS.[31] While controversy continues over whether this is a general approach applicable to all noncovalent complexes, there are documented cases in which information gained by gas-phase examination of protein–protein interactions appears to correlate well with data obtained from the same complexes in the liquid phase. An example is the enzyme ribonuclease S that requires a noncovalent association of a peptide and a protein for catalytic activity. Measurements of thermal denaturation of this complex in the gas phase and in solution indicated that the enthalpy of dissociation as determined by ESI-MS (gas phase) correlated well with measurements made in solution by calorimetry.[32] Other early studies showed a direct correlation between the individual observed relative abundances for a series of enzyme–inhibitor (E–I) complexes in the gas phase and their ranking or affinity found by calculating K_d values from traditional kinetic data.[33] The use of MS for the analysis of noncovalent protein complexes has been competently reviewed and is not further discussed in this paper.[31,34]

Other studies focused on the use of hydrogen–deuterium (H–D) exchange to examine higher order structural features of proteins by MS.[35] These experiments are based on the assumption that not all of the exchangeable hydrogens in a protein exchange at the same rate and that the rate of exchange is an indicator of structural properties of a protein. Examples of structural features that can be analyzed in this way include solvent accessibility, based on the observation that solvent-exposed hydrogens exchange more rapidly than those shielded from solvent access, and hydrogen bonding, based on the observation that hydrogens involved in hydrogen bonds exchange at a slower rate than those not involved in hydrogen bonds. With these concepts in mind, Anderegg and co-workers used ESI-MS to study the transition of a peptide from α-helical to a denatured configuration.[36,37] Even simpler experiments without H–D exchange have shown that ESI-MS can be used to monitor the transition of a protein in solution from native to denatured state. Such experiments rely on the empirical observation that ESI mass spectra of proteins known to be unfolded in solution indicate a higher charge state (a greater number of protons associated with the protein) than the identical protein not subject to denaturation.[38–41] Thus, the transition of a protein from a folded to a denatured state can be followed by ESI-MS by examining the charge state distribution of the protein molecular ions. Other, perhaps less controversial, applications of MS to study protein higher order structure include the identification of spatially juxtaposed amino acids by chemical cross-linking or the determination of the extent of heavy atom incorporation prior to X-ray diffraction of protein crystals.[42,43] The use of cross-linkers of a defined length that are chemically reactive to specific amino acid side chains such as the primary amine of lysine have been used to examine both intra- and inter-protein distances. A study of the yeast nuclear pore complex is a recent example of this approach.[44]

MS, in particular the application of ESI coupled on-line with high-performance separation techniques such as capillary electrophoresis (CE) and HPLC, has had a dramatic effect on the sensitivity and the speed with which the primary structure of proteins and peptides can be determined. Advances in separation techniques, particularly their implementation in miniaturized formats on-line with high-performance mass spectrometers,[45–50] and the development of miniaturized sprayers as ESI ion sources[51–53] have reduced the amount of peptide required for complete and routine sequence characterization from several picomoles of peptide[54,55] in the mid-1980s to a few femtomoles and below by the mid-1990s.[56–59] The development of mass spectrometric techniques of yet higher throughput and sensitivity is an essential component of the emerging field of proteomics and is still forcefully pursued today. Through incremental improvements in on-line separation methods, sensitivities in the sub-femtomole peptide detection and identification range have been achieved with commercially available ion trap mass spectrometers.[60,61] For simple mass measurement, sensitivities into the

Table 1. Sources for MS-Based Protein Identification Tools

sponsor (application)	uniform resource locator (URL)
Eidgenossische Technische Hochschule (MassSearch)	http://cbrg.inf.ethz.ch
European Molecular Biology Laboratory (PeptideSearch)	http://www.mann.embl-heidelberg.de
Swiss Institute of Bioinformatics (ExPASy)	http://www.expasy.ch/tools
Matrix Science (Mascot)	http://www.matrixscience.com
Rockefeller University (PepFrag, ProFound)	http://prowl.rockefeller.edu
Human Genome Research Center (MOWSE)	http://www.seqnet.dl.ac.uk
University of California (MS-Tag, MS-Fit, MS-Seq)	http://prospector.ucsf.edu
Institute for Systems Biology (COMET)	http://www.systemsbiology.org
University of Washington (SEQUEST)	http://thompson.mbt.washington.edu/sequest

zeptomole (10^{-21} mol) range have been observed with prototype FT-ICR-MS instruments.[62] The use of microfabricated devices connected on-line with ESI-MS[56,60,61,63] offers the exciting possibility of generating integrated analytical systems for sample manipulation, preparation, and analysis that promise to operate at unprecedented levels of automation, sensitivity, and throughput.

In this paper, we will review current methods that are the basis for proteome analysis, specifically mass spectrometric strategies for protein identification from biological matrixes, computational approaches for searching sequence databases, determination of protein expression levels, and characterization of phosphoproteins and phosphopeptides. We attempt to cover the most popular and promising techniques, but the review does not claim to be comprehensive. Furthermore, the mass spectrometers used for such studies, mainly the established MALDI time-of-flight[25] (TOF) mass spectrometer, the ESI-triple quadrupole[64,65] (TQ) mass spectrometer, the ESI ion trap[66,67] (IT) mass spectrometer, and the increasingly popular ESI-quadrupole-TOF [68–70] (Q-TOF), the ESI-TOF,[71] the MALDI-IT-TOF,[72] MALDI-TOF-TOF,[73] Fourier transform ion cyclotron resonance[74,75] (FT-ICR), MALDI-ion trap,[76,77] ESI-ion mobility,[78] and their respective operations, will not be described in detail here because this is the subject of another paper in this issue and other recent reviews.[79,80]

II. Methods for Protein Identification

Traditionally, proteins have been identified by de novo sequencing, most frequently by the automated, stepwise chemical degradation (Edman degradation) of proteins or isolated peptide fragments thereof.[81,82] These partial sequences were occasionally used to assemble the complete protein sequence from overlapping fragments but more frequently for the generation of probes for the isolation of the gene coding for the protein from a gene library. With the growing size of sequence databases, it became apparent that even relatively short and otherwise imperfect sequences (gaps, ambiguous residues) were useful for the identification of proteins. This was done by correlating information obtained experimentally from the analysis of peptides with sequence databases. The concept of identifying proteins by correlating information extracted from a protein or peptide with sequence databases rather than by de novo sequencing was significantly enhanced when it was realized that mass spectrometers were ideally suited to generate the required data. Furthermore, the methods initially developed for the isolation of small amounts of proteins and peptides for Edman sequencing were directly compatible with peptide analysis by LC-MS and LC-MS/MS. This further accelerated the implementation of mass spectrometric methods for protein identification.[21,82–88] Correlation of mass spectrometric data with sequence databases also depended on the development of novel search algorithms, a number of which are available on the worldwide web and listed in Table 1. Such algorithms use readily available constraints in a decision-making process that distinguishes the correct match from all other sequences in the database. The availability of complete sequence databases, the development of mass spectrometric methods, and the sequence database search algorithms therefore converged into a mature, robust, sensitive, and rapid technology that has dramatically advanced the ability to identify proteins and constitutes the basis of the emerging field of proteomics. In the following, we discuss the different approaches that have been developed for the identification of proteins by sequence database searching using data predominantly generated by MS.

A. Protein Identification Using Multiple Related Peptides

The methods described in this section use information obtained from analysis of multiple fragments of a single protein for database searching. Since the source of a specific fragment can only be unambiguously determined if a single, homogeneous protein is being analyzed, these methods require that proteins are highly purified. In the context of proteomics, a high degree of purification of multiple (ideally all) proteins in a sample is typically achieved by high-resolution two-dimensional gel electrophoresis (2DE). Therefore, these identification methods in conjunction with 2DE form the basis for many proteome projects.[89–92]

1. Nonmass Spectrometric Methods

As the subject of this paper is MS and proteomics, these methods are only mentioned peripherally. A more extensive treatment of nonmass spectrometric methods for use in proteome analysis can be found in a review by Wilkins et al.[18] It has been well-known for a long time that proteins differ considerably not only in their amino acid sequence but also in their amino acid composition.[93,94] Wilkins and co-workers therefore attempted to implement high throughput identification of proteins separated by 2DE by determining the accurate amino acid composition of

specific spots and submitting the data to a database search algorithm they developed.[95] It turned out that unambiguous protein identification by the amino acid composition alone was not always achieved and that secondary search constraints such as the isoelectric point and the molecular mass of the parent protein (as obtained from the spot coordinates in the 2D gel) were useful to increase the confidence of the search results and in many cases essential to obtain unambiguous identification. The method is quite sensitive to contaminating proteins present in the sample, either comigrating proteins or other contaminants, and has been essentially supplanted by mass spectrometric methods. Still the additional information provided by amino acid composition or even partial amino acid composition can be of value as an additional constraining parameter in sequence database searches. This was shown recently in a report that combined mass mapping (discussed below) with vapor-phase acid hydrolysis that was specific for protein cleavage after serine, threonine, aspartic acid, and glycine.[96]

2. Mass Spectrometric Methods

a. Principle of Peptide Mass Mapping. Peptide mass mapping is based on the insight that the accurate mass of a group of peptides derived from a protein by sequence-specific proteolysis (i.e., a mass map or fingerprint) is a highly effective means of protein identification. The principle behind protein identification by mass mapping is therefore quite simple conceptually and was implemented by several groups independently at approximately the same time.[85,97-100] Proteins of different amino acid sequence (Figure 1A) will, after proteolysis with a specific protease, produce groups of peptides the masses of which constitute mass fingerprints unique for a specific protein (Figure 1B). Therefore, if a sequence database containing the specific protein sequence is searched using selected masses (i.e., the observed peptide mass fingerprint), then the protein is expected to be correctly identified within the database. In the example shown in Figure 1B, the four peptide monoisotopic masses shown are sufficient to identify the protein as myoglobin and the species as *Equus caballus*. Various methods that automate this process have been developed and reviewed.[85] They vary in specific details but share the following sequence of steps: (i) Peptides are generated by digestion of the sample protein using sequence-specific cleavage reagents that allow residues at the carboxyl- or amino-terminus to be considered fixed for the search. For example, the enzyme trypsin that is popular for mass mapping leaves arginine (R) or lysine (K) at the carboxyl-terminus (Figure 1), and the N-termini of tryptic peptides (except for the N-terminal one) are expected to be the amino acid following a K or R residue in the protein sequence. (ii) Peptide masses are measured as accurately as possible in a mass spectrometer. An increase in mass accuracy will decrease the number of isobaric peptides for any given mass in a sequence database and therefore increase the stringency of the search. (iii) The proteins in the database are "digested" in silico using the rules that apply to the proteolytic method used

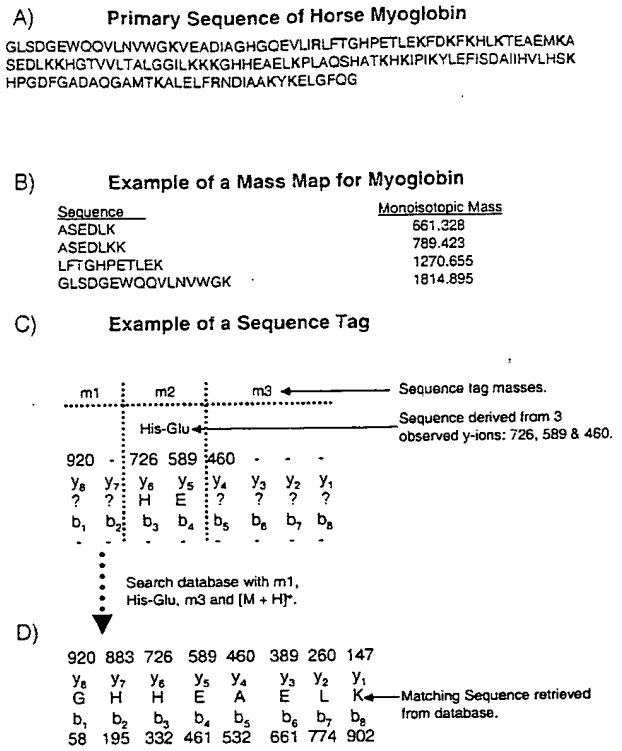

Figure 1. Protein identification. (A) Primary amino acid sequence of horse heart myoglobin with trypsin recognition sites (R and K) in bold. (B) Randomly selected monoisotopic mass values that serve as an example of a mass map. No simple rules exist for predicting all of the peptides that will be generated by digestion of a protein with trypsin. (C) One of the peptides from panel A presented as an example of the sequence tag method for protein identification. (D) Peptide fragment ion nomenclature and sequence of peptide from database determined with sequence tag information.

in the experiment to generate a list of theoretical masses that are compared to the set of measured masses. (iv) An algorithm is used to compare the set of measured peptide masses against those sets of masses predicted for each protein in the database and to assign a score to each match that ranks the quality of the matches. Obviously, for a protein to be identified its sequence has to exist in the sequence database being used for comparison. Both protein and DNA sequence databases are equally suited. If DNA sequence databases are being used, the DNA sequences are translated into protein sequences prior to digestion. The approach is therefore best suited for genetically well-characterized organisms where either the entire genome is known or extensive protein or cDNA sequence exists.

b. Pitfalls and Limitations of Mass Mapping. Protein identification by peptide mass mapping depends on the correlation of several peptide masses derived from the same protein with corresponding data calculated from the database. For this reason the method is suited neither for searches of EST databases nor for identification of proteins in complex mixtures if unseparated mixtures are proteolyzed. ESTs present a problem because they only represent a portion of a gene's coding sequence. Such segments may not be long enough to cover a sufficient number

of peptides observed in the mapping experiment to allow an unambiguous identification. Digests of unseparated protein mixtures present a problem for mass mapping because it is not apparent which peptides in the complex peptide mixture originate from the same protein. The mass mapping method is therefore most popular for the identification of proteins from microbial species for which complete genome sequences have been determined and for use with protein purification by 2DE where ancillary information on protein molecular weight and isoelectric point information can be used to aid identification. It is often combined with tandem MS of peptides (discussed later in this paper) in an iterative approach where as much information as possible is extracted by mass mapping, and this is followed by tandem MS to resolve the identification of any ambiguous remaining masses.

If a pure protein is digested and the resulting peptide masses are compared with the list of peptide masses predicted for that protein, two observations are typically made. First, not all of the predicted peptides are detected. Second, some of the measured peptide masses are not present in the list of masses predicted from the protein. The first problem, the missing masses, is usually due to a number of problems that can occur both before and during mass spectrometric analysis such as poor solubility, selective adsorption, ion suppression, selective ionization, very short peptide length, or other artifacts that cause sample loss or make specific peptides undetectable by MS. In rarer cases, e.g., in situations of alternative gene splicing, missing peptide masses may contain meaningful biological information. Unfortunately, it is not possible to distinguish between trivial and meaningful missing masses without further experimentation. Since a relatively low number of peptide masses are sufficient for the positive identification of a protein, missing peptide masses are not generally considered a problem. In contrast, unassigned peptide masses are a significant problem for protein identification by mass mapping and probably the single biggest source of misidentifications or missed identifications. Thus, to ensure that mass mapping results are reliable, it is important to understand the possible reasons for unassigned masses and to learn how to deal with them.[85,101-103] Unassigned masses may be observed for one or more of the following reasons: (i) Changes in the expected peptide masses by post-translational modification (e.g., phosphorylation adds a net 80 u to an amino acid mass), artifactual modifications arising from sample handling (such as oxidation of methionine), or post-translational processing (e.g., amino- or carboxyl-terminal processing). Some of these changes can be anticipated and incorporated into the search algorithm. (ii) Low fidelity proteolysis due to the presence of contaminating proteases that produce peptides unanticipated by the search algorithm (e.g., the presence of chymotryptic activity in a trypsin preparation) or missed cleavage sites. Again, this can be anticipated to some degree by the search algorithms. (iii) The presence of more than one protein in the sample. It needs to be stressed that bands in SDS gels frequently and spots in 2D gels occasionally contain more than one protein, even if the respective features appear concise and sharp. In some cases, additionally present proteins can be detected by iterative database searching with the masses left unassigned to the primary target protein. Keratins and other common proteins represent another source of protein contamination. (iv) The identified protein actually matches a sequence homologue or splice variant of that reported in the database. This must be confirmed using the sequence of genetically well-characterized species.[104] (v) The protein is misidentified (i.e., false-positive).

In this context, the specificity of the enzymes employed for protein digestion should be discussed in more detail. Obviously, the higher the fidelity of the enzyme in hydrolyzing peptide bonds, the more reliably the search can be done with a fixed amino- or carboxyl-terminus. Unfortunately, proteases are far from perfect enzymes. In addition to cleavage at expected amino acid residues, they tend to cleave at unexpected sites and tend to skip anticipated cleavage sites. The frequent observation that the protease products are not limited to the ones predicted from the expected enzymatic recognition sites is often due to contaminating protease activity but may also be due to a post-translational modification juxtaposed to the recognition site that blocks access by the enzyme. Furthermore, even highly purified trypsin appears to cleave at sites other than carboxyl-terminal to the expected recognition sites, K and R. The so-called "missed" cleavages produce "ragged" termini when two or more consecutive amino acids in a protein sequence are recognition sites for the enzyme. For example, when trypsin hydrolyzes myoglobin around the following sequence, ASEDLKKHGTVVTALGGILK (Figure 1A), it is expected that four peptides could be produced: ASEDLK, ASEDLKK, HGTVVTALGGILK, and KHGTVVTALGGILK. If this problem is anticipated, algorithms can be programmed to accommodate missed cleavages by allowing a given number to be entered as a parameter. Furthermore, the success of proteases to cleave proteins is dependent on accessibility to open stretches of primary amino acid sequence, and the native three-dimensional structure of the substrate protein will block access to many sites. Thus, proteins in solution are frequently not completely proteolyzed until they are denatured. Proteins separated by denaturing gel electrophoresis methods such sodium dodecyl sulfate−polyacrylamide gel electrophoresis (SDS−PAGE) are highly accessible to proteases and generally yield rich and reproducible peptide maps. In addition to the SDS that coats the protein and forces native interactions apart, reducing compounds such as dithiothreitol are used to chemically break disulfide bonds further opening the protein structure.

c. Secondary Parameters for Enhanced Peptide Mass Mapping. It should not be surprising that the single most critical experimental parameter for protein identification by mass mapping is the accuracy of the peptide mass measurement.[105-108] For any single measured peptide mass, the list of isobaric masses found in the database will decrease as the

mass accuracy of the measurement increases. Increased mass accuracy not only increases search speeds but also increases reliability in the score.[102,106,108,109] However, even in cases in which highly accurate mass measurements are possible, it is frequently desirable to use ancillary information to confirm true positives and to eliminate false-positive matches. Such information can be factored into the search algorithms to further constrain a search. The value of such ancillary information for peptide mass mapping has been extensively discussed,[105] and the power of this approach was illustrated in a recent report in which the presence of the relatively rare amino acid cysteine was used, along with other readily available constraints, to identify a protein within the yeast genomic sequence database from the accurate mass of a single cysteine-containing peptide.[110]

Good constraints should be easily obtainable experimentally and be highly discriminating. Quite a large number of viable experimental approaches for the generation of search constraints to enhance peptide mass mapping have been developed. They have the common objectives of deducing the presence of specific amino acids in the peptide analyzed and/or the location of a specific amino acid within the peptide without the use of tandem mass spectrometry. They differ in the way these objectives are pursued. The presence of specific amino acids in a peptide has been detected by site-specific chemical modifications. The number of modifications induced and therefore the number of the specifically targeted residues in a peptide is determined from the mass differential of a peptide before and after the modification reaction. Methyl esterification, which adds +14 u for each carboxyl group (i.e., side chains of aspartic acid, glutamic acid, or the carboxyl-terminus),[111] and iodination of tyrosine, which adds +126.9 u for each tyrosine,[112] are examples of such reactions. Similarly, the presence of cysteinyl residues was detected by isotopic labeling of sulfhydryl groups using a 1:1 mixture of acrylamide and deuterated acrylamide,[113] and the partial amino acid composition of peptides was determined by measuring the number of exchangeable hydrogens via hydrogen−deuterium exchange.[114] To locate specific residues within the peptide sequence, a single step of Edman degradation of unseparated peptide mixtures[115] or aminopeptidase[116] treatment have been used to identify the amino-terminal residue. Secondary proteolytic digestions (or a parallel digestions) using an enzyme with a different specificity from the one used for the primary digestion have been used to locate specific residues within a peptide.[111,114] The carboxyl-terminal residue(s) were identified by carboxypeptidase treatment of peptides or peptide mixtures.[116,117] In cases in which gel separated proteins are being identified, properties deduced from the position of the protein in the gel (protein mass for SDS−PAGE and protein mass and pI for 2DE) are also frequently used to confirm the identity protein analyzed.

d. Generation of Data for Peptide Mass Mapping. Data for use with peptide mass mapping are commonly obtained via MALDI-TOF analysis. However, any mass spectrometer capable of generating mass accuracies around 100 ppm or better at 1000 u (i.e., 100 parts in 1000.000 or in the example accurate to the first place past the decimal point), in particular ESI-TOF and FT-ICR instruments, can be used to generate a mass map. For MALDI, analytes are spotted onto a metal plate either one at a time or, in a higher throughput format, multiple samples on the same plate. The samples are usually tryptic digests from proteins separated by 2DE, although proteins purified by other separation methods are also compatible with the method. Before deposition of the analytes, the matrix is placed on the plate or mixed in with the sample. The matrix will absorb energy from the laser causing the analytes to be ionized by MALDI. The m/z ratio of the ions is then typically measured based on the flight time in a field-free drift tube (as opposed to ion mobility MS where a field pushes ions through a gas) that constitutes the heart of the time-of-flight mass (TOF) analyzer. Using internal calibration on monoisotopic masses, a mass accuracy of 5 ppm at 1000 u can be achieved. An additional bonus for samples isolated from biological sources is that MALDI is compatible with biological buffers such as phosphate and Tris and low concentrations of urea, nonionic detergents, and some alkali metal salts. Peptide m/z ratios are calculated based on the energy equation ion $E = 1/2mv^2$ that accounts for contributions from kinetic energy, mass, and velocity. At a constant energy, low molecular weight ions will travel faster than high molecular weight ions—flight times of ions are inversely proportional to the square root of their molecular mass.

An inherent problem with the MALDI process is the small spread of kinetic energy that occurs during ionization. The spread reduces the resolving power and prevents the observation of the natural isotope distribution, even of small peptides. Two approaches, an ion mirror (reflectron) and "time-lag focusing" (a.k.a. delayed extraction), have been implemented in commercial instruments to overcome this problem. A reflectron is a device located at the end of the flight tube opposite from the ion source that decelerates the ions and then re-accelerates them back out of the reflectron toward a second detector. This is achieved by applying a decelerating voltage that is slightly higher than the accelerating voltage at the source. It has been observed that ions of lower kinetic energy do not penetrate as far into the reflectron as those of higher energy. Consequently, deeper penetrating high-energy ions can catch up, thereby decreasing the initial energy spread. The second approach to correct the initial spread of kinetic energies during MALDI is the time-lag focusing technique initially developed by Wiley and McLaren in 1953 and more recently reintroduced as "delayed extraction".[118,119] In this method, the MALDI ions are created in a field-free region and allowed to spread out before the extraction voltage is applied to accelerate them for their flight through the drift tube. This results in a significantly decreased energy spread of ions and thus higher resolution. Delayed extraction also limits peak broadening due to metastable decomposition from ions colliding in the source during continuous ion extrac-

tion. The effects of these improvements are significant. Delayed extraction can increase the mass resolution to ~2000−4000 for peptides in a linear instrument and, if combined with a reflectron instrument resolution, can further increase to ~3000−6000.[119,120]

e. Examples of Proteome Projects by Peptide Mass Mapping. *Haemophilus influenzae* has been the subject of one of the most extensive proteome efforts based on 2DE and mass mapping to date.[92] Soluble proteins were separated by 2DE. To enhance the separation range, immobilized pH gradient strips of various pH ranges[121] and second dimension SDS gels with different acrylamide concentrations and electrophoresis buffers with different trailing ions were used. Low-copy-number proteins were visualized by employing a series of protein extraction and chromatographic steps that included heparin chromatography, chromatofocusing, and hydrophobic interaction chromatography. Cell envelope-bound proteins were separated electrophoretically, either by immobilized pH gradient strips or a two-detergent system in which a cationic detergent was used in the first dimension and an anionic detergent in the second. A combination of MALDI-TOF MS and amino acid composition analysis identified a total of 502 proteins out of a genome of approximately 1742 ORFs.

In contrast or in addition to cataloguing the proteins expressed in a particular cell or tissue, global protein expression profiles, if analyzed quantitatively, can be useful to differentiate cell types or the same cell type in different physiological or pathological states. A recent study used 2DE to better examine the taxonomic relationship between several divergent yeast species. Yeast strains used in the brewing industry are known to be hybrid strains of at least two different genomes.[92] By analyzing the proteins expressed by three commonly used brewing strains by 2DE and comparing the resulting patterns to the patterns and identified protein spots for *S. cerevisiae*, it was established that *S. carlsbergensis*, *S. monacensis*. and *S. pastorianus* represented two divergent evolutionary patterns. One pattern originated from an *S. cerevisiae*-like genome and the other from *S. pastorianus* strain NRRL Y-1551. Numerous additional proteome projects have been attempted or are in progress (see current information at www.expasy.ch). Generally, peptide mass mapping is chosen as the method of choice for protein identification if a complete genomic sequence database is available. Protein identification using tandem mass spectrometry (see below) has proven to be superior however for the identification of proteins from species with large and incompletely sequenced genomes.

B. Protein Identification Using Single Peptides

Different amino acid compositions and permutations of an amino acid sequence can result in isobaric peptides. The amino acid sequence of a peptide is therefore more constraining than its mass for protein identification by sequence database searching.[122] At the mass accuracy achieved with the MALDI-TOF mass spectrometers that are frequently used for peptide mass measurement (10−100 ppm), several peptide masses from the same protein are required for unambiguous identification, whereas the amino acid sequence of even a relatively small peptide can uniquely identify a protein. The number of proteins that are targeted for identification in typical proteomics projects by far exceed the capacity of the traditional chemical sequencing methods, and mass spectrometric methods for the rapid generation of amino acid sequence information from intact proteins, while exciting, still require further development.[122−126] Large-scale protein identification therefore critically depends on tandem mass spectrometry for the generation of sequence-specific spectra for peptides. In this section, we describe the methods used for the generation of sequence-specific peptide mass spectra and their use for large-scale protein identification. These methods were initially designed to identify purified proteins (typically homogeneous protein spots in 2D gels) faster, more sensitively, and more reliably. More recently, essentially the same methods have been applied to also identify proteins in complex mixtures.[127]

1. Protein Identification via Sequence-Specific Peptide Mass Spectra

Tandem mass spectrometers and, to a more limited extent, single-stage mass spectrometers have the ability to fragment peptide ions and to record the resulting fragment ion spectra. For tandem mass spectrometers such as triple quadrupole, ion trap, or quadrupole/TOF instruments, fragment ion spectra are generated by a process called collision-induced dissociation (CID) in which the peptide ion to be analyzed is isolated and fragmented in a collision cell, and the fragment ion spectrum is recorded. Typically, but not exclusively, these types of mass spectrometers are used in conjunction with ESI. For the most part, the low-energy CID spectra of peptides generated by ESI-MS/MS are of high quality and are sequence specific. Other mass spectrometric methods including fragmentation of high-energy ions and post-source decay (PSD) in a MALDI MS also produce sequence-specific fragment ion spectra.[128] Generally, these are more difficult to interpret than the low-energy CID spectra generated by ESI-MS/MS and are not further discussed in this paper.

a. Background to Peptide Fragmentation. Tandem mass spectra generated by the fragmentation of peptide ions in the gas phase at low collision energy are dominated by fragment ions resulting from cleavage at the amide bonds. Very little amino acid side chain fragmentation is observed. Such spectra are much less complex than the high collision energy spectra generated in magnetic sector or TOF/TOF instruments. The low-energy CID spectra generated by the types of mass spectrometers most frequently used in proteomics are therefore relatively simple to interpret, and a straightforward nomenclature for annotating the MS spectra has been adapted (Figure 2). The nomenclature differentiates fragment ions according to the amide bond that fragments and the end of the peptide that retains a charge after fragmentation.[129,130] If the positive charge

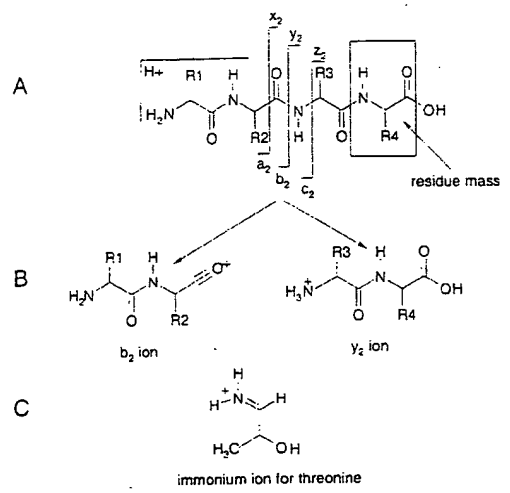

Figure 2. Peptide fragment ion nomenclature. (A) Nomenclature for peptide fragment ions that form via cleavage of bonds along the peptide backbone. (B) Example structure for *b* and *y* ions. Note that the b_2 ion specifically is thought to take the form of a cyclic oxazole[131] rather than a highly unstable acyl cation as shown. (C) Immonium ion for threonine drawn for simplicity. Not all amino acids generate immonium ions and then not to the same extent.

associated with the parent peptide ion remains on the amino-terminal side of the fragmented amide bond, then this fragment ion is referred to as a *b* ion. However, the fragment ion is referred to as a *y* ion if the charge remains on the carboxyl-terminal side of the broken amide bond. Since in principle every peptide bond can fragment to generate a *b* or *y* ion, respectively, subscripts are used to designate the specific amide bond that was fragmented to generate the observed fragment ions. *b* ions are designated by a subscript that reflects the number of amino acid residues present on the fragment ion counted from the amino-terminus, whereas the subscript of *y* ions indicates the number of amino acids present, counting from the carboxyl-terminus. These individual fragment ion *m/z* values as shown in Figure 1C can be easily calculated from the amino acid sequence, using the nominal (i.e., monoisotopic value rounded to an integer value) residue masses found in Table 2. To calculate the masses of the *b* ion series (Figure 1C), 1 u (for 1 H) is added to the nominal mass for the first residue. In the example indicated, the nominal mass for glycine (nominal mass = 58) is added to indicate the mass of the b_1 ion. To calculate the masses for the b_2, b_3, and following fragment ions, this process is continued by the addition of the nominal mass for the second, third, and following amino acid residues, respectively, until the final, carboxy-terminal amino acid is included. The *b* ion series will stop at 902 or 18 u short of the $[M + H]^+$ mass. To calculate the masses for the *y* ion series (Figure 1C), 19 u (for H_3O^+) is added to the nominal residue of the carboxy-terminal amino acid. In the example indicated, this residue is lysine with a mass of 147 u. As for the *b* ion series, this process is continued with the addition of the nominal mass of the following amino acids until the $[M + H]^+$ value is reached. While it is relatively simple to calculate the elements of the *b* and *y* ion series from the

Table 2. Residue and Immonium Ion Masses of 20 Common Amino Acids

amino acid (3/1 letter codes)	nominal residue mass	immonium ion mass
alanine (Ala /A)	71	44
arginine (Arg/R)	156	129
aspartic acid (Asp/ D	115	87
asparagine (Asn/N)	114	88
cysteine (Cys/C)	103	76
glutamic acid (Glu/E)	129	102
glutamine (Gln/Q)	128	101
glycine (Gly/G)	57	30
histidine (His/H)	137	110
isoleucine (Ile/I)	113	86
leucine (Leu/L)	113	86
lysine (Lys/K)	128	101
methionine (Met/M)	131	104
phenylalanine (Phe/F)	147	120
proline (Pro/P)	97	70
serine (Ser/S)	87	60
threonine (Thr/T)	101	74
tryptophan (Trp/W)	186	159
tyrosine (Tyr/Y)	163	136
valine (Val/V)	99	72

peptide sequence, it is much less straightforward to read the amino acid sequence from the CID spectrum of a peptide ion. This is mainly because peptide fragmentation under the conditions encountered in the collision cell of a mass spectrometer are sequence dependent, and the rules for fragmentation are not completely understood. Additionally, the drawing of the *b* ion structure as an acyl cation as presented in Figure 2 is historical. It has been recently shown that the b_2 ion specifically is much more likely to exist as a cyclic oxazole ring[131] rather than the acyl cation[54] as shown.

b. Properties of Peptide Fragment Ion Spectra. The CID spectrum of a peptide ion acquired at low collision energy can be considered a composite of many discrete fragmentation events. Each peptide tandem mass spectrum will contain *b* and *y* ions as well as other fragment ions that can be used to interpret the amino acid sequence. These include diagnostic ions generated by the neutral loss of specific groups from amino acid side chains (e.g., the loss of ammonia (−17 u) from Gln, Lys, and Arg or of water (−18 u) from Ser, Thr, Asp and Glu) and low mass ions that result from the fragmentation of amino acids down to a basic unit consisting of the side chain residue and an immonium functionality (Figure 2). The *b* ion series also often shows a satellite ion series in which each signal is 28 u lower than the corresponding *b* ion. These signals result from the neutral loss of carbon monoxide and are referred to as an *a* ion series. CID spectra can be further complicated by the presence of internal fragment ions that represent some contiguous sequence of amino acids in the peptide. These are generated if a specific peptide ion undergoes two or more fragmentation events. Empirical observation shows that internal fragments often occur if either proline[132,133] or aspartic acid[134] residues are present in a sequence and even more so at any aspartyl-proline bond,[135] indicating that not all peptide bonds have the same propensity to fragment during low-energy CID. For the same reason, the relative

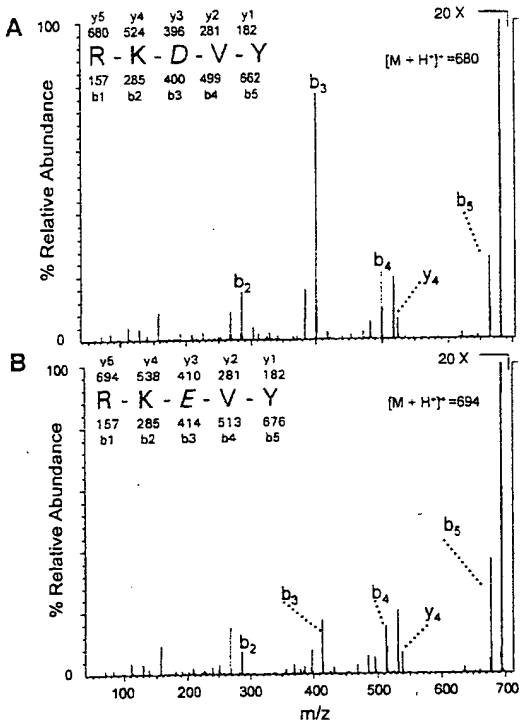

Figure 3. Tendency of peptides to fragment at aspartic acid. The $[M + H]^+$ ions for peptides, identical in sequence except for a substitution of aspartic acid (A) with glutamic acid (B), were subjected to CID under identical conditions using electrospray ionization on a triple quadrupole mass spectrometer.

intensity of fragment ions in peptide CID spectra is uneven and somewhat unpredictable. Some of the rules that control peptide ion fragmentation in a collision cell have been determined;[54,131,136] many others remain to be studied. If a proline residue is present in a peptide sequence, the most intense ions in the CID spectrum will generally be due to fragmentation on the amino-terminal side of proline.[132] This is thought to occur because the gas-phase basicity of the proline imide bond is greater than that for any of the amide bonds. Under the moving proton hypothesis for CID, the proton available for fragmentation is therefore statistically more likely to be at this imide bond as compared to the amide bonds in the peptide.[136] Additionally, it is known that peptides that contain aspartic acid tend to fragment at the carboxyl-terminal adjacent amide bond (i.e., Asp-Xxx; Figure 3A). This observation may be due to the ability of the aspartic acid side chain to influence the gas-phase basicity of the adjacent carboxyl-terminal amide bond via formation of a transient six-membered ring between the carboxylic acid group of the Asp side chain and the nitrogen of the adjacent amide bond. Leading credence to this is the observation that when glutamic acid, chemically similar to aspartic acid in that it contains one extra methylene carbon giving the side chain more degrees of freedom, is substituted for aspartic acid into a peptide, the Glu-Xxx bond (Figure 3B) does not fragment as readily as the Asp-Xxx bond (Figure 3A).[134]

Thus, the quality of peptide tandem mass spectra is dependent on the sequence location of amino acids, amino acid side chain basicity, amino acid side chain structure, and charge state of the peptide ion fragmented. If proteins are completely digested with trypsin, then lysine or arginine residues will be present at the carboxyl-terminus of all peptides except for the C-terminal peptide of the original protein. A charge sequestered by lysine or arginine at the C-terminus tends to produce a more complete series of *y* ion fragments than will be generated by peptides produced by protein digestion with chymotrypsin or other protease where lysine and arginine are distributed throughout the sequences rather than at the C-terminus. Additionally, $[M + 2H]^{2+}$ ions of peptides will produce tandem mass spectra of higher quality than those from either $[M + H]^+$ or $[M + 3H]^{3+}$ peptide ions. The $[M + 2H]^{2+}$ peptide ions fragmented under low-energy CID produce spectra, although there are exceptions such as when proline and/or histidine are internal to the peptide sequence, that contain $[M + H]^+$ fragment ions that are more readily interpreted than tandem mass spectra of $[M + 3H]^{3+}$ and higher charge states that produce multiply charged fragment ions.

c. Generation of Tandem Mass Spectra. Peptide fragmentation for the purpose of protein identification, either for single isolated proteins or on a proteome wide scale, is most often carried out by CID in a triple quadrupole (TQ),[47] ion-trap (IT)[58] or quadrupole time-of-flight (QTOF)[69,70] mass spectrometer and to a lesser extent by PSD-MALDI-MS.[128,137] Among these methods, fragmentation by PSD-MALDI-MS is least well-controlled partly because only a few parameters of the experiment can be readily varied (i.e., laser energy and type of matrix). Furthermore, peptide ionization by MALDI generates mostly $[M + H]^+$ ions that do not produce readily interpretable tandem mass spectra. PSD-MALDI-MS therefore has a lower success rate, and the spectra are in many cases of lower quality than the spectra of the same peptides produced by ESI and low-energy CID in a collision cell. For these reasons, CID of peptides in TQ, IT, and QTOF mass spectrometers is more frequently used for protein identification.

If one or a few peptides derived from a pure protein are being analyzed by CID, the time required for the mass spectrometer to select a specific peptide ion for CID, to fragment the ion, and to record the fragment ion spectrum is fast in comparison with the time required to prepare the sample and to analyze the data, even if computer algorithms are used for data interpretation. The identification of one or a few proteins is therefore frequently carried out in a manual mode in which the *m/z* ratio of the peptide ion selected for CID is controlled by the operator of the instrument. Manual precursor ion selection and control of CID conditions has the advantage that the fragmentation conditions can be optimized by an experienced operator for each precursor ion during the experiment. A particularly successful implementation of this approach uses a variation in ESI called nanospray in which a peptide sample is introduced at very low flow rates, typically nanoliters per minute, into the mass spectrometer.[53,138-140] The low sample consumption afforded by the nanospray tech-

nique allows for extended observation and accumulation of the ion signals and generally yields CID spectra of excellent quality.

Proteomics requires the analysis of large numbers of proteins, each one potentially generating multiple peptide fragments. The time required for the analysis of a single peptide by CID therefore rapidly becomes limiting in proteome studies if each ion has to be manually identified and selected. Therefore, protocols for automated, instrument-controlled selection of precursor ions have been developed. In these methods, ion selection for CID is under computer control and based on signals observed in the full-scan mass spectrum (i.e., data-dependent MS/MS).[58,141-145] In the most basic implementation of the method, the system selects the most intense ion (i.e., base peak) in a given *m/z* range for CID, carries out CID on that ion, and then writes that parent ion *m/z* value to a list for dynamic exclusion from further CID for some defined time. This iteration begins again as the next most intense ion in the original mass spectrum is chosen for CID. If the sample is introduced by infusion of an unseparated peptide mixture from a syringe or by nanospray, the system will walk through all ions above a preset threshold. If the sample is introduced from an on-line separation method such as capillary electrophoresis or HPLC, the observed mass spectrum will change continuously during the separation. If complex peptide mixtures are separated and analyzed on-line by ESI-MS/MS, it is frequently the case that a larger number of peptide ions are detected in a chromatographic peak than can be subjected to CID during the time available. Therefore, even with dynamic exclusion to limit redundant CID of the most abundant peptides, not all of the peptide ions with a signal intensity above the preset selection threshold will be analyzed. The consequence of this situation is an apparent compression of the dynamic range of the mass spectrometer. Some of the more sophisticated systems alleviate this problem by modulating the flow rate of the HPLC [58,142] or capillary electrophoresis[141] separation system into the ESI source in a signal-dependent manner. Generally, the flow rate is reduced as long as ion signals exceeding the preset threshold are detected in a chromatographic peak and re-accelerated between peaks. These "peak parking" procedures take advantage of the concentration-dependent nature of ESI[146] and allow tandem MS spectra to be acquired on some of the lower abundance species that might normally be missed.

For peptide mass mapping, the information collectively contained in the masses of several peptides derived from the same protein is used for protein identification by database searching. In contrast, the CID spectrum of a single peptide can, in principle, contain a sufficient amount of information for unambiguous identification of a protein.[147,148] Therefore, if a mixture of several proteins is concurrently digested, the components of the mixture can be identified based on the CID spectra, provided that at least one CID spectrum per protein is generated. It is therefore no longer necessary to separate proteins to homogeneity prior to proteolysis. The composition of even relatively complex protein mixtures can now be ascertained without purification of individual proteins by using two-dimensional chromatography methods coupled to data-dependent ESI-MS/MS.[127]

While most of the discussion has been about low-energy CID, it is worth noting a new type of sequencing mass spectrometer that provides CID at high energy using a TOF/TOF mass spectrometer.[73] This was achieved by coupling two TOF mass spectrometers together via a collision cell between them. The new design combines the advantages of MALDI such as high sensitivity for peptide analysis, relative insensitivity to salts, surfactants, and other contaminants with high-energy CID where amino acids such as isoleucine and leucine can be distinguished by side-chain fragmentation. As with other types of sequencing mass spectrometers, a complete CID spectrum can be acquired in a single acquisition, obviating the need to sum as many as 10 spectra as is necessary with PSD on a single TOF mass spectrometer. Additionally, the MALDI-TOF/TOF mass spectrometer promises to be capable of acquiring tandem mass spectra at a rate that is an order of magnitude above the capabilities of IT and QTOF instruments, which will be significant for proteome studies of mammals where the number of ORFs is 10−100 times that of microbials.

d. Protein Identification Using Tandem Mass Spectra. Regardless of the method of fragmentation, low-energy spectra will contain redundant pieces of information such as overlapping *b* and *y* series ions, multiple internal ions from the same peptide, and immonium ions. This redundancy makes fragment ion spectra an extremely rich source of sequence-specific information, but it also complicates the interpretation of the sequence. This is illustrated by the CID spectrum shown in Figure 4A that represents a low-energy spectrum of the $[M + 2H]^{2+}$ ion of angiotensin. Even with the peptide sequence available to aid spectral interpretation, there are only a few ions that can be easily assigned unambiguously without considering the possibility of doubly charged fragment ions. Furthermore, it is not immediately apparent to which ion series a particular ion belongs (i.e., is it a *y* or *b* ion?); that is to say that directionality is not apparent a priori from the fragment ions. While the *b* ion series will often be accompanied by an *a* ion series, for a complete unknown even this clue may not be of much help because not every *b* ion will have an associated *a* ion. In the case illustrated, the abundance of doubly charged ions arise from the amino acids arginine, proline, and histidine that can serve as sites for proton sequestration. An additional complication arises from the fact that this peptide was not produced by cleavage of a protein with trypsin. Therefore, the peptide does not contain a carboxyl-terminal sequestered charge that helps reveal the *y* ion series and, conversely and even worse for interpretation, it contains an internal arginine. Compare the spectrum in Figure 4A to the one in Figure 4B, which is from the $[M + 2H]^{2+}$ ion of peptide resulting from autolysis of trypsin. This tryptic autocatalytic peptide has no charged residues

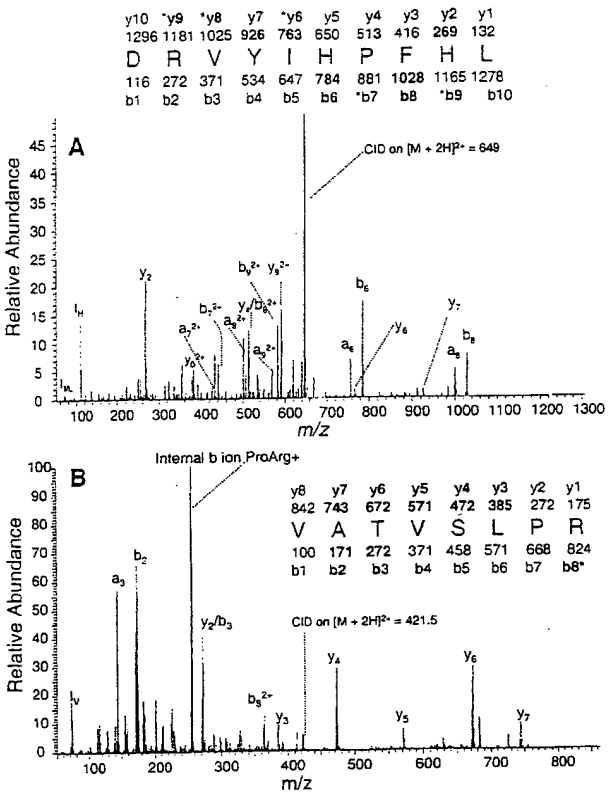

Figure 4. Annotated tandem mass spectra. The annotated tandem mass spectra for (A) $[M + 2H]^{2+}$ of angiotensin and (B) a peptide from the protease trypsin commonly produced by autocatalysis of trypsin. Both spectra were obtained at low collision energy using electrospray ionization on a triple quadrupole mass spectrometer.

located in the interior of the sequence and has arginine at the carboxyl-terminus. Both of these factors are advantages that help one to distinguish the *y* ion series over the *b* ion series.

From this example, it is apparent that the manual, explicit interpretation of sequence information contained in a CID spectrum is complex, labor- and decision-intensive, and slow. Clearly, the manual interpretation of CID spectra would constitute a significant bottleneck in every proteome project based on protein identification by CID spectra. Since at least some of the rules that determine the fragmentation of peptide ions in a low-energy collision cell have been known, it seemed sensible to devise computer algorithms that interpret or at least assist an operator in the interpretation of the CID spectra. While such attempts at computerized de novo sequence interpretation have been progressing, the most significant advances in rapid protein identification using CID spectra came from the development of algorithms that correlate peptide CID spectra with sequence databases, automatically and without the need for user input.

2. De Novo Peptide Sequencing

The basic problem presented by the de novo interpretation of a CID spectrum of a peptide with unknown sequence is one of directionality. It is a priori impossible to determine which ions belong to the *b* ion and which ones belong to the *y* ion series; without this assignment, the sequence cannot be interpreted. Therefore, a number of methods have been developed that address this problem. They can be categorized into MS- and MS/MS-based methods. The MS-based methods attempt to derive sequence information without tandem MS. They are based on the generation of peptide ladders in which individual elements of the ladder differ in length by one amino acid. The mass differences between the elements of the ladder as determined by MS therefore indicate the amino acid sequence of the peptide analyzed. In the MS/MS-based approaches, CID spectra are generated as described above. To identify the *b* and *y* ion series, respectively, and therefore the directionality of the peptide sequence, a number of chemical modification procedures have been introduced.

Peptide ladders for sequence analysis without using tandem MS can be generated by chemical or enzymatic degradation of peptides and can proceed from either the amino- or the carboxyl-terminus. The chemical methods for determining amino-terminal sequences employ the principles of the Edman degradation chemistry and therefore require a free primary amine at the amino-terminus. Two variations of the method have been described. In the first, Edman reagents that react with the amino-terminal primary amine are added at the beginning of each sequencing cycle. A "ladder" of ions reflecting the peptide sequence is generated by blocking the amino-terminus of a small fraction of the peptides in each cycle via the addition of a small amount of isocyanate blocking agent to the Edman reagent. The blocked peptides are not degraded during subsequent cycles. After a number of degradation cycles, the sample is recovered and analyzed by MALDI-MS. The sample is a mixture of peptide fragments (i.e., a peptide ladder), the masses of which differ by the mass of the amino acid residue cleaved off in each sequential cycle.[149,150] The second approach, referred to as "nested peptide sequencing", is based on the addition of an aliquot of the peptide/protein at each cycle of the Edman degradation process. In this method, the Edman chemistry process is driven to completion using an excess of a volatile reagent that can be removed in each cycle by evaporation. Just as with ladder sequencing, a peptide sequence is generated when masses of products are measured.[151] Both methods require a free amino-terminus, and neither can resolve the isobaric residues isoleucine and leucine. However, glutamine (residue mass 128.13) and lysine (residue mass 128.17), difficult to distinguish in CID spectra, can be easily distinguished because the ϵ-amino group of lysine side is modified with the Edman reagent. In principle, both methods can cope with mixtures of peptides whose intact masses are sufficiently different to avoid overlap of the degradation products.[150] Alternatives to the chemical stepwise degradation peptide sequence ladders have also been generated by truncation of peptides by amino and carboxyl peptidases. This approach has the advantage that reactions can be conducted with very small quantities of starting material and can be carried out directly on the

MALDI probe surface. The enzymatic reactions are stopped by the addition of a matrix,[115-117] and the method is easily amenable to time course studies. Neither the chemical nor the enzymatic methods to generate sequence ladders for peptide sequencing by MS are routinely used.

If high-quality CID spectra can be obtained by MS/MS (i.e., spectra containing complete *b* and *y* ion series), it may be possible in some cases to determine de novo the peptide sequence.[152] Unfortunately, interpretation of CID spectra is often a difficult process even with high-quality data. Confidence in the final sequence assignment may remain low unless a biological assay is available to test the proposed sequence. It is much more common for CID to be of marginal quality due to poor ion statistics arising from either too little peptide or a peptide chemistry that is partially refractory to fragmentation. The difficulty of de novo sequence analysis, and in particular of identifying the respective ion series, has led to several ingenious methods that help elucidate the sequence. For instance, fragment ions belonging to a *y* ion series can be identified by conducting proteolysis with trypsin in a buffer containing 50% v/v $H_2{}^{18}O$/50% v/v $H_2{}^{16}O$. This takes advantage of the hydrolytic action of trypsin that adds a molecule of water across each amide bond that is hydrolyzed. Every peptide with the exception of the peptide derived from the carboxyl-terminus will appear as doublets differing in mass by 2 u.[153-155] If both isotopically labeled parent ions are selected for CID together, only fragment ions with an intact carboxyl-terminus will appear as doublets separated by 2 u. While the method simplifies the fragment ion assignment, it also necessarily reduces the signal intensity of each *y* ion, therefore making sequencing at very high sensitivities more difficult. The use of high-resolution MS instruments such as the QTOF tandem mass spectrometer for analysis of isotopically labeled *y* ions as described above can result in complete interpretation of peptide spectra in a single experiment.[156] The method has also been successfully applied with an ion-trap mass spectrometer.[157] An alternative chemical approach that again distinguishes *y* ions from *b* ions involves methyl esterification of the carboxyl groups in a peptide. This reaction increases the mass of the peptide by 14 u for each carboxyl group (unmodified carboxyl-terminus, side chains of aspartic acid, glutamic acid).[54] If there are no acidic residues in the peptide and only the carboxyl-terminus is esterified, the mass of each signal in the *y* series will be increased by 14 u as compared to the corresponding signals obtained from the unmodified peptide. To distinguish the *y* series ions, this approach therefore requires a comparison of the fragment ion spectra obtained from the methylated peptide and the original, underivatized peptide. Methods to specifically tag the amino-terminal residue with the purpose of identifying the *b* ion series have also been utilized. In these experiments, peptide amino groups were derivatized with a reagent containing a permanent positive charge in the gas.[54,111,158] Additionally, a method that takes advantage of the moving proton hypothesis for fragmentation of tryptic peptides was described that derivatizes the amino-terminus with an acidic reagent.[159] This approach helps achieve a charge balance between the basic carboxyl-terminal residues of either lysine or arginine (in tryptic peptides) and the moving proton available for fragmentation.

With the development of computer algorithms and large-scale databases, the need for de novo sequencing of peptides is clearly declining. However, for the sequence analysis of the many proteins from species for which no genomic or expressed sequence tag database is available, the art of reading out amino acid sequences from CID spectra will remain important. An excellent, detailed tutorial was recently published.[160]

3. Manual Generation of Peptide Sequence Tags

In an attempt to accelerate protein identification using CID spectra of peptides and to take advantage of the extensive sequence databases available, two basic types of approaches have been employed. They have in common that they correlate the information in the fragment ion spectra with sequence databases using computer algorithms. They differ in the way the information is extracted from the CID spectra. The first relies on a partial manual interpretation of the spectrum to identify consecutive elements of a particular (*b* or *y*) ion series to provide a partial sequence (i.e., any contiguous set of *b* or *y* ions). This partial sequence is then used together with the mass of the parent ion mass to determine by subtraction the mass difference between the parent ion mass and the total mass of the amino acids that constitute the partial sequence tag. This calculation provides a mass the sum of which constitutes a possible amino acid composition. If the protein was digested with trypsin, then the amino acid in the carboxyl-terminal position can be guessed as either arginine or lysine. Together all of this information provides a peptide-sequence tag that can be used for searching databases.[161,162] Using the example in Figure 1C, a partial internal sequence tag, His-Glu, may be derived from observing three *y* ions at *m/z* 726, 589, and 460. This along with the parent peptide mass, 920 *m/z*, provides residue masses with unknown sequence at the amino- and carboxyl-terminus of 194 and 441, respectively. This information is then used as input for an algorithm to search a database for protein identification. More recently, the generation of sequence tags has been automated, further increasing the utility of this approach.[163] A recently published algorithm, SHERENGA, intended to automate the interpretation of CID spectra for de novo sequencing currently falls short of this ambitious goal but is very useful for the generation of sequence tags or for validating sequence database matches generated by automated database searching tools (see below).[164] SHERENGA automatically learns fragment ion types (i.e., *b* vs *y* ion directionality) and intensity thresholds from a database of spectra generated from peptides of known sequence. It can be applied to data from any type of mass spectrometer provided that the spectra in the database are generated by the same type of instrument as the spectra to be analyzed.

4. Automated Interpretation of CID Spectra

The second approach for computer-assisted protein identification via CID spectra uses the uninterpreted fragment ion pattern and mass of the parent ion as input for sequence database searching. This approach is exemplified by the algorithm SEQUEST.[165] The algorithm first creates a list of peptide masses isobaric to the observed mass on which CID was carried out by searching the database of choice for possible amino acid sequences that can generate peptide masses to match the mass of the parent peptide. For each of these candidate peptides, the program calculates the masses of the fragment ions expected, without consideration of chemical information that is reflected in the relative intensities of fragment ions (i.e., all predicted fragment ions are equal in intensity) and generates a theoretical CID mass spectrum for comparison. The program then compares the observed fragment ion spectrum with the top 500 theoretical fragment ion spectra using cross-correlation algorithms. Each comparison then receives a score that is ranked relative to all other possibilities according to a number of parameters such as the number of fragment ions predicted versus found. Once an answer is arrived at, then it is important to confirm this top score using a functional assay if possible and, at the very least, to manually check the predicted sequence.[147,166] The constraints on database searching of a given stretch of peptide sequence are so powerful that the tandem MS spectrum of a single peptide can be adequate for protein identification in an EST database.[147] The approach is easily automated[167] and can also be adapted to find peptides carrying specified post-translational modifications by instructing the program to anticipate modification at specific residues (e.g., 80 u is added to phosphorylated residues such as serine, threonine, or tyrosine).[167,168] A list of some Internet sites with protein identification resources developed by these and other investigators can be found in Table 1.

5. Accurate Mass Tags

A type of instrument that is gaining in popularity for proteomic analysis is FT-ICR-MS. The advantages of FT-ICR-MS are severalfold and include high resolution over a broad *m/z* range,[74] high sensitivity/dynamic range,[62] and high mass accuracy.[109,110,169] Both MALDI and ESI-FT-ICR MS are proving to be an order of magnitude more sensitive for peptide detection than standard triple quadrupole and ion-trap technology. As little as tens of attomoles of peptide loaded onto a MALDI probe can be detected.[170] Additionally for ESI, 10 amol of phosphoangiotensin II loaded on a 50 μm i.d. capillary column and eluted via a standard acetonitrile gradient was used to measure masses and fragment peptides via ESI-FT-MS.[110,171] The ability to measure peptide masses to 0.1 ppm at 1000 u creates the possibility of using the mass of a single peptide as a unique identifier for a protein when working with a specific genome of relatively small size such as *S. cerevisiae* (~6000 ORFs) or *H. influenza* (~1700 ORFs). Whereas the peptide mass normally only provides a subset of possible amino acid compositions[122] that could be combined to define the measured mass, very high mass accuracy allows the use of accurate mass tags (AMTs) for protein identification in small genomes but probably will not be adequate for larger mammalian genomes such as humans, which may have anywhere from 40 000 to 100 000 ORFs (as of 2000, the number of human ORFs is disputed).[109] Furthermore, in cases where the mass measured to an accuracy of 0.1 ppm is not unique, it may be possible to use readily available constraints such as the presence of cysteine[110] in a peptide to positively identify the parent protein. The advantage of protein identification from the accurate mass of single peptides circumvents the most significant disadvantage of data-dependent MS/MS, which is that it is impossible (no matter how many tricks are tried) to perform tandem MS on every ion presented in a chromatographic window in time.[58,141,143] Additionally, while a specific peptide ion is selected for CID, other coeluting ions are not selected for tandem MS and are therefore excluded for analysis. These excluded ions may or may not be selected for tandem MS in subsequent iterations of the process. By circumventing the need for data-dependent tandem MS, all ions present in a chromatographic window in time are used for protein identification. This means that low abundance proteins that are passed over by the data-dependent methods because of low signal intensity can be identified by taking an accurate mass snapshot of each chromatographic window.[110]

C. Protein Identification in Complex Mixtures

The field of proteomics has grown out of the mature technology of high-resolution two-dimensional gel electrophoresis (2DE) for protein separation and quantitation[172,173] and the increasingly refined technologies described above for the identification of separated proteins.[85] Today, 2DE and protein MS represent an integrated technology by which several thousand protein species can be separated, detected, and quantified in a single operation, and hundreds of the detected proteins can be identified in a highly automated fashion by sequential analysis of the peptide mixtures generated by digestion of individual gel spots. It is commonly assumed that 2DE-MS is a suitable technology base for global proteome analysis based on its ability to display, quantify, and identify thousands of proteins in a single gel.[91] However, closer examination of the proteins routinely identified by proteome studies suggests that 2DE-MS does not represent a truly global technique. Specific classes of proteins are known to be absent or underrepresented in 2D gel patterns. These include very acidic or basic proteins, excessively large or small proteins, and membrane proteins. In addition, by examining codon bias values of proteins identified from 2D gels, it has now been shown that the 2DE-MS approach is incapable of detecting low abundance proteins without pre-gel enrichment.[174] Codon bias is a measure of the propensity of an organism to selectively utilize certain codons, which result in the incorporation of the same amino acid residue in a growing

polypeptide chain. It is also thought to be a good measure of protein abundance because highly expressed proteins generally have large codon bias values.[175] It has been previously shown that almost without exception proteins identified from 2DE-MS experiments of whole yeast lysates are abundant proteins (codon bias values >0.2).[91,174,176,177] This is a striking finding since more than one-half of all yeast genes have codon bias values <0.1, thus making these proteins undetectable by 2DE without prior enrichment. Furthermore, accurate quantitation of proteins separated by 2DE is complex and of limited dynamic range, particularly if high-sensitivity staining methods are being used.[178] Clearly, the detection and quantification of low abundance proteins such as trnscription factors, protein kinases, and other proteins of regulatory function is an important component of proteomics and incompatible with the standard 2DE-MS approach.

These limitations inherent in the 2DE-MS or 2DE-MS/MS approach to proteomics and the emerging ability to identify the components in protein mixtures using data-dependent, automated LC-MS/MS and sequence database searching have catalyzed the development of new, chromatography-based methods for the identification of the proteins contained in complex mixtures without the need for separation of the mixture into the individual protein components.[127,144,179-181] This is accomplished by the digestion of the unseparated proteins and the analysis of the resulting complex peptide mixture by LC-MS/MS. What is most exciting about these types of experiments is the sheer numbers of peptides that can be sequenced automatically in reasonable time frame in a single analysis. Using an ion trap mass spectrometer, an MS strategy was reported for highly complex peptide mixtures that employs a single MS scan followed by five MS/MS (sequencing) scans on the five most-intense peptide ions in that scan. Up to 10^4 sequencing attempts were recorded in a single LC-MS analysis of 60-min duration.[143] To increase even further the number of peptides that can be sequenced in a single analysis, techniques where complex peptide mixtures are separated on-line by a combination of cation-exchange and reverse-phase chromatography have been reported.[127] The success of these methods to identify and catalog rapidly and reliably large numbers of proteins in complex mixtures makes them powerful tools for descriptive proteomics.

For proteomics in general, the separation sciences continue to make great strides in analyzing complex mixtures and offer the potential for circumventing gel electrophoresis as a preparative tool for MS. Over the past decade, gel electrophoresis followed by proteolysis of individual stained protein bands has been the most common method for separating proteins prior to MS identification.[182] However, a number of laboratories have been investigating the use of chromatography only based approaches that bypass the electrophoretic based preparative gel methods altogether, except for diagnostic purposes. Two-dimensional or orthogonal chromatography approaches such as cation exchange followed by reverse-phase on-line with tandem MS have been successfully used to identify proteins in complex mixtures after proteolysis.[127,183,184] Even more complex approaches have used computer-controlled setup with an autosampler, five columns, and three 10-port switching valves to allow a series of steps to be performed on-line, obviating the need for any manual transfers of materials. The strategy included the following: immunoaffinity chromatography, desalting and buffer exchange on a mixed-bed strong ion-exchange absorbent, enzymatic digestion on an immobilized trypsin column, capture of peptides on a short perfusion capillary reversed-phase column, and final separation on an analytical reversed-phase column with on-line MS/MS analysis.[185] These orthogonal chromatography techniques have as a common goal the circumvention of the weakness of data-dependent analysis of complex mixtures, namely, that very complex mixtures of peptides exceed the capacity of these computer routines to carry out CID on all of the peptides present in a given full-scan mass spectrum. Thus, by fractionating complex peptide mixtures on-line, these two-dimensional chromatography methods help extend the dynamic range of the overall analysis.

The reduction of the complexity of ions presented during any chromatographic window in time for the data-dependent MS routines that automatically select ions for fragmentation is an important focus in the further development of protein mixture analysis by LC-MS/MS or LC/LC-MS/MS. This is being approached by chemistry-based reduction of the peptide sample complexity and by the development of mass spectrometers that can more completely sample the peptides present in a sample. Selective tagging of specific, relatively rare functional groups in peptides has been introduced as a strategy to reduce the complexity of peptide samples. Reports to date used relatively specific alkylation reactions to selectively tag the sulfhydryl side chains in cysteine-containing peptides, but any other specific reaction targeting a rare group in a peptide would be equally suitable.[110,182,186] If MALDI-TOF/TOF instruments,[73] capable of conducting tandem MS at rates an order of magnitude faster than currently possible with IT or QTOF instruments, are combined with existing separation systems to fractionate peptides separated by HPLC or ion exchange directly onto a MALDI target, then more complete sequence coverage for any given sample should be one obvious result. This new design of sequencing mass spectrometer selects ions for CID with a TOF mass spectrometer and then analyzes fragment ions in a second TOF mass spectrometer rather than post-source decay.[187] Combining with pre-fractionation from HPLC separations would allow one fraction to be analyzed repeatedly and perhaps completely, thus obviating the need to repeat a chromatographic separation. Furthermore, it is possible that for proteomic studies where genomes are completely sequenced that developments in FT-ICR-MS may circumvent the need for "serial" tandem MS to identify proteins. It has recently been demonstrated that simultaneous or "parallel" trapping of a group of unrelated peptides (i.e., not derived from the same parent protein) followed by collective fragmen-

tation of all peptides in the ICR cell can be used to identify all proteins from which each peptide was derived without true tandem MS (i.e., selection of a single ion followed by fragmentation).[188] In this case, high mass accuracy provided a critical benefit in that it allowed simultaneous identification of multiple and different proteins using only a single peptide ion mass determined prior to fragmentation to 1 ppm together with a single peptide fragment ion mass also determined to 1 ppm. It is at least hypothetically possible that a similar approach could be implemented at lower mass accuracy on any TOF mass detector using in-source CID to collectively fragment peptides.

D. Analysis of Protein Expression

A promising and exciting new use for MS in proteomics involves not just the identification of proteins but also the determination of protein expression levels (relative quantity) between two different pools of proteins. Obtaining expression data for proteins as is routinely done for mRNA is important because protein expression levels often are diagnostic of a given cellular state and are not directly related to levels of mRNA expression.[174] Methods exist for the global comparison of mRNA as a function of cellular state,[189] and they are widely used to identify clusters of genes for which expression is idiotypic of a cellular state. Until recently, no such methods with adequate dynamic range existed for obtaining relative protein expression measurements for most or all the proteins expressed by a cell or tissue (see discussion related to dynamic range of 2DE). To add a quantitative dimension to non-2DE-based proteome analyses, the venerable technique of stable isotope labeling has been adapted for protein analysis.[190] The method involves the addition to a sample of chemically identical but stable isotope (e.g., ^{2}H, ^{13}C, ^{15}N, etc.) labeled internal standards. In the case of ESI and MALDI, ionization efficiency can be quite variable for peptides of different sequence or even identical peptides from different MALDI spots or for ESI different HPLC conditions. Thus, the best internal standard for a candidate peptide is a peptide of identical sequence but labeled with stable isotopes. Quantitative protein profiling is therefore accomplished when a protein mixture (reference sample) is compared to a second sample containing the same proteins at different abundances and labeled with heavy stable isotopes (Figure 5). In theory, all the peptides in the sample then exist in analyte pairs of identical sequence but different mass. As the peptide pairs have the same physicochemical properties, they are expected to behave identically during isolation, separation, and ionization. Thus, the ratio of intensities of the lower and higher mass components provides an accurate measure of the relative abundance of the peptides (and hence the protein) in the original protein mixtures. Three groups initially and independently reported measuring quantitative protein profiles based on stable isotopes.[186,191,192] The techniques differ in the method of incorporation of heavy isotopes (i.e., in vivo (Figure 5A) or in vitro (Figure 5B)) and in the analytical procedures used to measure protein expression and identify proteins. Chait and co-workers grew one yeast culture on medium containing the natural abundance of the isotopes of nitrogen (^{14}N, 99.6%; ^{15}N, 0.4%) while another culture was grown on the same medium enriched in ^{15}N (>96%).[191] After an appropriate growing period, the cell pools were combined, and proteins of interest were extracted and separated by RP-HPLC and then by SDS–PAGE. In-gel digestion of excised spots of interest resulted in peptide fragments, which were identified by peptide mass mapping. Each ^{15}N incorporated shifted the mass of any given peptide upward, leading to a paired peak for each peptide. The percent error of the experimental technique was found to be excellent (±10%). Smith and co-workers used stable isotope-depleted media to impart a specific isotope signature into proteins.[192] They compared the cadmium stress response in *Escherichia coli* grown in normal and rare isotope-depleted (^{13}C-, ^{15}N- and ^{2}H-depleted) media. Intact protein mass measurements were carried out by FT-ICR-MS. While no protein was positively identified, the expression ratios for 200 different proteins were compared. Clearly, stable isotope metabolic protein labeling using ^{15}N-enriched or depleted media permits quantitative protein profiling in conjunction with either

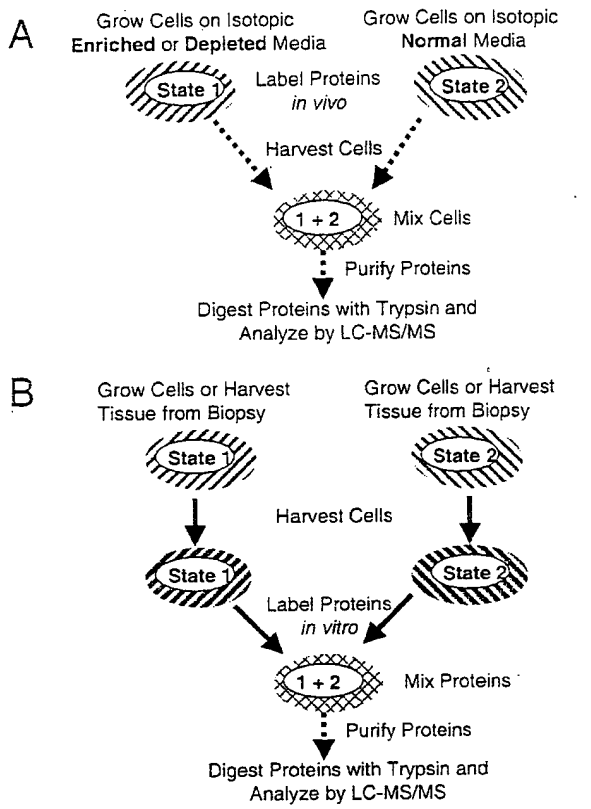

Figure 5. Scheme for determining protein expression. In the absence of a protein array on a chip as currently exists for measuring mRNA expression via cDNA arrays, MS as a universal detector can be used to determine relative protein expression between two samples of interest. (A) In vivo and (B) in vitro isotopic labeling of proteins followed by protein purification, proteolysis, and analysis by LC-MS/MS to determine both the identity of proteins and relative ratio of proteins expressed.

2DE or other separation techniques. However, this method has several disadvantages. First, the method does not allow for the analysis of proteins directly from tissue. Second, the stable isotope-enriched media are costly and may themselves affect cellular growth and protein production. Third, the increase in nominal mass due to stable isotope incorporation is not known until the sequence is determined. Therefore, protein identification has to precede quantitation. We have recently published a novel method for quantitative protein profiling based on isotope-coded affinity tags (ICAT).[186] In this method, the stable isotopes are incorporated post-isolation by selective alkylation of cysteines with either a heavy (d8) or normal (d0) reagent. The two protein mixtures are then mixed. At this point, any optional fractionation technique can be performed to enrich for low abundance proteins or to reduce the complexity of the mixture, while the relative quantities are strictly maintained. Prior to analysis, the protein mixture is digested with trypsin and passed over a monomeric avidin−agarose column. Because the ICAT label contains the stable isotope information as well as a biotin tag, ICAT-labeled (cysteine-containing) peptides are selectively isolated for analysis by microcapillary LC-ESI-MS/MS. The ratio of ion intensities from coeluting ICAT-labeled pairs permits the quantification while a subsequent MS/MS scan provides the protein identification. The advantages of the ICAT strategy are severalfold. First, the method is compatible with any amount of protein harvested from bodily fluids, cells, or tissues under any growth conditions. Second, the alkylation reaction is highly specific and occurs in the presence of salts, detergents, and stabilizers (e.g., SDS, urea, guanidine hydrochloride). Third, the complexity of the peptide mixture is reduced by isolating only cysteine-containing peptides. Fourth, the ICAT strategy permits almost any type of biochemical, immunological, or physical fractionation that makes it compatible with the analysis of low abundance proteins. There are two disadvantages to the method. First, the size of the ICAT label (~500 Da) is a relatively large modification that remains on each peptide throughout the MS analysis. This can complicate the database searching algorithms, especially for small peptides (<7 amino acids). Second, the method fails for proteins that contain no cysteines. However, only a small percentage of proteins are cysteine-free (8% in yeast), and ICAT reagents with specificities to groups other than thiols could be synthesized.

A third approach to determine protein expression levels that does not involve stable isotopic labeling is possible, but it requires that the sample number is high enough for statistical significance to be proven.[193,194] In this case, normalization between cellular states relies on proving that one MS signal is statistically higher or lower in ion current than the signal for the same peptide from a different sample. Obviously, these studies are inherently difficult to conduct given the "relative" nature of MS detectors. However, MS has the advantage of being a universal detector and will continue to play an important role in measuring protein expression until a protein chip or other device is designed that can fill this role. Thus, regardless of the method, quantitative analysis of proteomes promises to provide a complimentary technique to mRNA expression for developing clinical diagnostics and studying basic genetics.

III. Proteomes and Post-Translational Modifications

A. Proteomes

A proteome is neither a static entity nor the product of the direct translation of gene sequences into protein sequences. In the previous sections, we discussed MS-based methods for the identification of proteins in complex samples and for monitoring quantitative changes in their abundance. In this section, we will discuss the challenges posed by the diverse post-translational mechanisms that process and modify proteomes permanently or reversibly and current methods used for the analysis of the products of these mechanisms. MS is an essential component of virtually every current strategy but is by itself insufficient to analyze post-translational modifications and processing.

1. The Analytical Challenge

If all the relevant properties of proteins were apparent from the gene sequence and could therefore be precisely predicted, the justification for establishing complex and expensive platforms for proteome analysis would be low. Proteome analysis is based on the expectation that the information gained by direct protein analysis exceeds or complements that obtained by the more readily available methods for gene sequence analysis. In addition to the sequence and abundance, the properties of proteins that are of particular interest to biologists include their subcellular location and state of modification, their function and state of activity, and the nature of interacting proteins. It is not obvious how these diverse properties can be determined systematically and quantitatively for a single protein. Extending these measurements to a proteome wide scale is even more challenging and, despite recent advances, remains largely unachievable with current methods. In this context, applications of MS have mostly focused on the characterization of protein−protein complexes and the analysis of post-translational modifications. These two topics will be further discussed in this paper. It can also be anticipated that in the near future emerging technologies will be combined with MS to extend the range of proteome-wide analyses that indicate the functional state of a biological system. These include the ability to measure quantitatively the specific activity of specific classes of enzymes in complex protein samples[195] and methods such as laser capture microdissection[196,197] and advanced methods for subcellular fractionation and the isolation of cellular organelles[198] to determine the subcellular location of proteins and the dynamics of protein trafficking patterns. We expect that the development of proteomics technologies related to the

analysis of post-translational processing and control will be very exciting and fruitful in the next few years.

2. Analysis of Protein–Protein Complexes

Most cellular functions are not performed by individual proteins but rather by protein assemblies also termed multi-protein complexes. It is therefore frequently assumed that proteins that specifically interact also partake in the same function, and the identification of specifically interacting proteins is an important component of the proteomics quest to study the function of biological processes. In general, the methods described above for the analysis of protein mixtures in general are also suited for the analysis of protein complexes, and some of the most scientifically rewarding applications of protein MS have involved the analysis of protein complexes. Yates and co-workers[127] identified more than 70 proteins present in the yeast ribosome in a single analysis using LC/LC-MS/MS. Peptide mass mapping was used to exhaustively analyzed the composition, architecture, and transport mechanism of the yeast nuclear pore complex.[199] Chemical cross-linking and MS were used to examine the spatial organization of multi-protein complexes,[44] and the components of the T-cell receptor complex[200] have been study by SDS–PAGE and tandem mass spectrometry. Such projects critically depend on the ability to isolate the target complex cleanly and in good yields. To this end, a tandem affinity purification (TAP) method has been developed and demonstrated its effectiveness by examining the yeast spliceosome.[201,202] In eukaryotes, seven Sm proteins bind to the U1, U2, U4, and U5 spliceosomal snRNAs while seven Sm-like proteins (Lsm2p–Lsm8p) are associated with U6 snRNA. Another yeast Sm-like protein, Lsm1p, does not interact with U6 snRNA. Using the tandem affinity purification (TAP) method, Lsm1p was identified among the subunits associated with Lsm3p. Coprecipitation, using antibody (so-called immunoprecipitation) experiments, demonstrated that Lsm1p together with Lsm2p–Lsm7p formed a new seven-subunit complex. The two related Sm-like protein complexes were purified, and the proteins recovered were identified by MS. MS and the TAP purification scheme were thus used to confirm the association of the Lsm2p-Lsm8p complex with U6 snRNA. Approaches such as this can be used in conjunction with and to test results from purely genetic methods such as the yeast two-hybrid approach.[203] Coprecipitation methods combined with mass spectrometric identification of proteins complement the genetic approaches such as the TAP purification scheme and the popular yeast two-hybrid system, which can be set up in a high throughput format but suffers from a lack of specificity. The above discussion of just one simple pathway reveals the complexity involved in trying to understand protein–protein interaction pathways on a global scale. In the coming years, proteome laboratories will attempt to map out all known pathways in select cellular systems using MS, genetics, and molecular biology.

B. Post-Translational Modifications

1. Background

MS is the method of choice for the detection and identification of post-translational modifications (PTMs). In principle, the methods used for protein identification are also applicable to the analysis of PTMs. For a number of reasons, PTM analysis is however significantly more complex than simple protein identification: (i) Proteins are frequently modified to a low stoichiometry only. Therefore, a high sensitivity of detection for the modified peptides is required. (ii) While proteins can be identified by the sequence or the CID spectrum of a single peptide, the identification of PTMs requires the isolation and analysis of the specific peptide that contains the modified residue(s). (iii) The bond between the PTM and the peptide is frequently labile. It may therefore be difficult to find conditions that maintain the peptide in its modified state during sample work up and ionization. (iv) More than 200 different types of protein modifications have been described.[204] The total sequence space containing all the potential modified protein sequences is therefore enormous. Likewise, the space required to comprehensively treat the procedures for the analysis of all types of post-translational modifications by far exceeds the space available for this chapter. We focus this discussion on protein phosphorylation because it is both biologically important and illustrates the challenges posed by PTM analysis. Many of the general approaches and specific methods discussed in the context of protein phosphorylation are however directly or with minor adaptations applicable to the analysis of different types of modifications.

a. Protein Phosphorylation: Catalysis and Biology. Among the large number of PTMs described to date, only a few have been shown to be reversible and thus potentially of regulatory importance in biological processes. Of these, protein phosphorylation has received the most attention and is the best understood with respect to both the enzymes involved catalyzing the phosphorylation/dephosphorylation reactions and the functional consequences of protein phosphorylation. The most common type of protein phosphorylation studied involves the formation of phosphate esters bonds with the hydroxyl side chains of serine, threonine, and tyrosine. Two counteracting enzyme systems, kinases and phosphatases, catalyze protein phosphorylation and dephosphorylation, respectively. The structures, specificities, and regulation of the most common of these is well-studied and reviewed.[205–207] There are assumed to be hundreds of protein kinases/phosphatases differing in their substrate specificities, kinetic properties, tissue distribution, and association with regulatory pathways. For example, analysis of the complete genomic DNA sequence from *S. cerevisiae* for sequence motifs that are thought to be indicative of protein kinases (and phosphatases) predicts 123 different protein kinases (and 40 protein phosphatases) that could be expressed. Thus, approximately 2% of expressed yeast proteins are involved in protein phosphorylation reactions, and presumably a much larger number of

proteins are phosphorylated under specific physiological conditions. In addition to phosphate esters involving the side chains of the hydroxyl amino acids, phosphoramidates of arginine, histidine, and lysine and acyl derivatives of aspartic and glutamic acid have also been observed. Some of these modifications are not typically observed unless specific precautions are taken to prevent their elimination during protein isolation.[208] For example, phosphohistidine is a relatively common modification, at least in prokaryotes, but it is completely eliminated by the acidic conditions commonly used for protein staining in polyacrylamide gels.

Protein phosphorylation generally exerts it regulatory function by altering the structure and thus the function of target proteins.[207,208-210] The functional consequences of site-specific protein phosphorylation are dramatic and diverse. Glycogen phosphorylase is the prototypical enzyme that is controlled by phosphorylation.[206] A single phosphorylation event involving a serine residue close to the N-terminal induces a long-range allosteric change and thus a change in the catalytic activity.[211] Work on the protein tyrosine kinase p56lck demonstrated that phosphorylation at a single serine, possibly by a mitogen-activated protein (MAP) kinase,[212] can alter the substrate specificity of an enzyme.[210] Tyrosine phosphate-induced protein interactions involving SH2 domains[213] have developed a common mechanism for stabilizing protein complexes that perform complex biological functions, particularly in intracellular signal transduction.[214-217] Protein phosphorylation has also been shown to be involved in targeting the protein IκB for ubiquitin-mediated destruction[218] to prolong the half-life of the yeast protein SST-2[219] and to control the DNA binding and transactivating activities of transcriptional activators.[220] This list of biological functions of protein phosphorylation is by no means complete, and additional functions controlled by protein phosphorylation are likely to yet be discovered. However, this overview illustrates some of the main challenges of proteome-wide phosphorylation studies, such as the large number of proteins phosphorylated, the diverse physiological conditions that induce the phosphorylation of specific groups of proteins in a cell, and the determination of the functional significance of an observed phosphorylation event.

b. Process of Protein Phosphorylation Analysis. The principal aims of essentially any protein phosphorylation study are 3-fold (Figure 6). The first is to determine the amino acid residues that are phosphorylated in vivo in a protein present in a cell in a given biological state. The second is to identify the kinase/phosphatase(s) responsible for catalyzing the specific reaction. The third is to understand the functional significance of the observed phosphorylation events for the biology of the cell. Among these aims, only the first can be directly addressed by MS and will be covered further. It is apparent, however, that the proteomics technologies discussed in the first part of this paper and other systematic or global research methods are playing an increasingly significant role in the comprehensive analysis of the

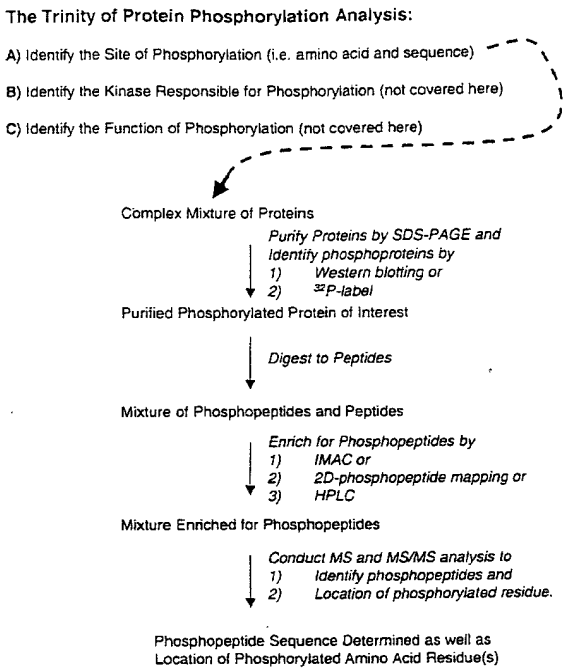

Figure 6. Phosphoprotein analysis scheme. Examples of some of the many possible routes for identifying phosphorylated amino acids in a protein and the local amino acid sequence around the phosphorylated residue.

biology of protein phosphorylation. For example, the systematic, quantitative comparison of the abundance of a large number (ideally all) of proteins expressed in a wild-type cell and in a cell expressing a regulatory protein with the site(s) of phosphorylation mutated might shed light on the processes that are controlled by the regulatory protein in question. Likewise, the availability of all the 123 currently recognized protein kinases of the S. *cerevisiae* proteome in the form of GST-fusion proteins overexpressed one at a time in genetically engineered yeast strains should make the identification of the kinase(s) that can potentially phosphorylate a particular site in a target protein relatively straightforward.[221]

MS would most efficiently identify the precise residues phosphorylated in a proteome if phosphorylated peptides could be selectively isolated from a digest of the proteins contained in a sample, separated, and analyzed by tandem mass spectrometry. Although some progress toward a method[222,223] with these properties has been achieved, no proteome-wide phosphorylation studies have been reported to date. Essentially all the experiments reported to date that involve the identification of phosphorylation sites first attempt to purify the phosphorylated protein to homogeneity before the protein is enzymatically fragmented and the phosphopeptides are isolated and mass spectrometrically analyzed. Since proteins are frequently phosphorylated to low stoichiometry (i.e., only a small fraction of a given protein may be phosphorylated) and at multiple sites (giving rise to differentially phosphorylated forms of the same protein), it is frequently difficult to isolate quantities of in vivo phosphorylated proteins that are sufficient for analysis by even the most sensitive MS methods.

Many protein phosphorylation studies are therefore performed with proteins modified by in vitro kinase reactions that can generally be scaled up to produce larger amounts of the phosphorylated protein. However, before sites of phosphorylation determined by in vitro studies can be accepted as biologically significant, the occurrence of the same phosphorylation sites in vivo needs to be established. This is frequently accomplished by comparing two-dimensional phosphopeptide (2DPP) maps of the same protein phosphorylated in vivo and in vitro.[224] Described in greater detail below, the technique essentially allows confirmation that the easily produced in vitro phosphoprotein can be used as a surrogate for the in vivo protein. Comigration of phosphopeptides generated by in vivo and in vitro phosphorylation, respectively, of the same protein is then taken as an indication that the same site is phosphorylated in vivo and in vitro[215,224] and that the in vitro system can be used as a source for the generation of peptides for mass spectrometric analysis.

Whether the site of phosphorylation is determined directly from a protein phosphorylated in vivo or a protein phosphorylated by an in vitro kinase reaction, the actual determination of the phosphorylated residue(s) generally consists of the following steps: (i) detection and purification of the phosphoprotein, (ii) enzymatic or chemical cleavage of the phosphoprotein into peptides, (iii) isolation of the phosphopeptides from nonphosphorylated peptides or at least phosphopeptide enrichment, and (iv) characterization of the phosphopeptides by MS. These steps are described in greater detail below.

2. Detection and Purification of Phosphoproteins

If a total cell lysate or any other complex protein mixture is analyzed, it is not a priori apparent which proteins, if any, are phosphorylated. A protein may first be suspected of being phosphorylated due to the presence of a protein band that migrates slightly slower in SDS−PAGE than the protein being studied. However, observation of two closely migrating bands in a one-dimensional gel or the observation of an array of spots of similar molecular mass but different isoelectric points in a two-dimensional polyacrylamide gel is insufficient to identify a particular protein as a phosphoprotein. It simply provides a suggestion that phosphorylation is a possible cause for the change in electrophoretic mobility. Stronger evidence that the slower migrating band may contain a phosphorylated form of the protein may be obtained if before electrophoresis the protein mixture was subjected to in vivo or in vitro labeling with ^{32}P followed by detection of the bands containing ^{32}P labeled proteins by autoradiography or storage phosphorimaging. Metabolic or in vivo radiolabeling is accomplished by incubating cells or tissue with $^{32}PO_4$ for a period long enough to equilibrate the cellular ATP pool with ^{32}P so that protein kinases can phosphorylate substrates. Protein phosphorylation in vitro is generally performed by kinase reactions using [γ-^{32}P]ATP as the source of the radiolabel and crude fractionated cell lysates or purified kinases as the source of the kinase activity. If labeling of the proteins with radioisotopes is not an option, the use of antibodies to detect phosphorylated proteins after electrophoretic separation (i.e., Western blotting) is a possible alternative method. This nonradioactive approach has been most successful for detection of tyrosine phosphorylated proteins.[225] A panel of tyrosine phosphate specific and very sensitive antibodies (4g10, py20) have been developed that have been invaluable tools for studying tyrosine phosphorylation.[226] The development of antibodies with specificity for phosphoserine or phosphothreonine[227,228] has been less successful, and reports using these reagents for the detection of proteins phosphorylated at these sites have been less frequent.[229] If sufficient amounts of the phosphoprotein are present in the gel for mass spectrometric detection, the bands can be excised, digested with trypsin, and subjected to data-dependent tandem MS as described above. We have recently described simple guidelines to estimate the amount of phosphoprotein present in a sample and to assess the chances for success of mass spectrometric analysis.[230] Most commonly, the phosphorylated peptides represent minor components in the peptide sample and need to be enriched prior to MS analysis in order to raise their signal above the level of the general background of low-intensity ions. This is especially true when data-dependent MS protocols are used that select ions for CID based on their relative signal intensities in the MS scan or based on a predetermined signal threshold.

3. Phosphopeptide Separation Methods

Among the common separation techniques, two-dimensional (electrophoresis/TLC) phosphopeptide mapping on cellulose plates (2DPP),[230] RP-HPLC,[231] 1D and 2D high-resolution gel electrophoresis,[232] immobilized metal affinity chromatography (IMAC),[233] and capillary electrophoresis[224] have been successfully used for the separation of phosphopeptides. Peptide separation techniques help to concentrate phosphopeptides and therefore increase the signal-to-noise ratio. Reproducible patterns of separated, radiolabeled phosphopeptides can also be used to quantitatively determine changes in the phosphorylation state of a protein as a function of time or cellular state, provided that a quantitative method is available for their detection. Peptide separation methods also effectively remove nonpeptidic contaminants, thus facilitating the detection and analysis of low-abundance phosphopeptides. Each one of the five methods (2DPP mapping, RP-HPLC, high-resolution gel electrophoresis, IMAC, and capillary electrophoresis) is compatible with further mass spectrometric analysis of the separated peptides. Finally, each of the separation methods can be used to calculate an absolute quantity of purified phosphopeptides if they are radiolabeled by in vitro kinase reactions to a known specific activity.[230] Knowing the amount of purified phosphopeptide is critical for the choice of a suitable MS strategy and for an assessment of the chances for success of the experiment.

a. Two-Dimensional Phosphopeptide Mapping. In 2DPP mapping, peptides are separated in a first dimension by electrophoresis on a thin-layer

cellulose plate and in the second dimension by thin-layer chromatography (TLC) on the same plate.[234] Separated, ^{32}P-radiolabeled phosphopeptides are then detected by autoradiography or storage phosphor imaging. The method provides important qualitative and quantitative data about the phosphorylation state of the protein not available from any other method. These include the following: (i) the maximum number of phosphorylation sites as maps generally produce more spots than there are phosphorylation sites because of differential processing by proteases. (ii) The relative stoichiometry of phosphorylation among all phosphopeptides is provided by autoradiographic intensity. (iii) The relative hydropathy of the separated phosphopeptides is apparent from the tangential separations of electrophoresis and TLC. A significant advantage of 2DPP mapping is that it produces purified phosphopeptides that can be analyzed, after extraction from the plate, directly by MS methods.[235] If 2DPP mapping is intended as a preparative method for MS/MS analysis, then the amount of protease added should be kept as low as is practical. Use of excess protease will be obvious when analyzing a "purified" phosphopeptide by MS because autocatalytic products from the protease will dominate the spectra and may prevent analysis of the phosphopeptide. Furthermore, the 2DPP method is very sensitive and can be even more sensitive than MS methods because detection is by integration of radioactive decay over potentially very long time periods. Finally, the high degree of pattern reproducibility achieved by the method makes 2DPP the method of choice for projects in which the state of phosphorylation of a protein under different conditions needs to be analyzed using, for example, time courses and different induced states of activation.

b. High-Resolution Gel Electrophoresis. Preparative methods using 1D or 2D electrophoresis to purify phosphopeptides on polyacrylamide gels, respectively, were recently published.[232] In the 2D method, nondenaturing gel isoelectric focusing was combined with alkaline 40% polyacrylamide gel electrophoresis for phosphopeptide separation and comparative pattern analysis as is done with 2DPP mapping. Phosphopeptides were detected by autoradiography or storage phosphorimaging of ^{32}P-labeled samples. Edman sequencing rather than MS was used to identify the proteins and to determine the sites of phosphorylation, but the method is presumably adaptable to MS-based methods. The method promises the same results as 2DPP mapping except that the recovered samples might be less contaminated with the non-peptide components that compete with peptide analytes for ionization in the mass spectrometer and that are carried along with the peptides after extraction from the cellulose matrix used for 2DPP mapping. Unlike with 2DPP mapping, there is the potential for loss of specific phosphopeptides if electrophoresis is not closely monitored. Regardless, the method is appealing because it uses relatively common equipment whereas 2DPP mapping requires purchase of specialized equipment for conducting the first dimension electrophoretic separation.

c. Ion Metal Affinity Chromatography. One commonly overlooked difficulty with phosphopeptide analysis is the low stoichiometry of phosphorylation. In such cases phosphopeptide(s) are present in the sample in very small amounts as compared to the nonphosphorylated peptide with the same sequence and the other peptides derived from the digested protein. It is thus difficult to identify phosphopeptides by MS techniques even though their presence in the sample is confirmed by ^{32}P label that was detected in a 2DPP spot, a whole protein digest, or a HPLC fraction. As mentioned before, data-dependent tandem MS methods often fail to identify minor species in a sample because priority for selection for CID goes to the most intense ion detected in the MS scan. To alleviate this problem, selective enrichment of phosphopeptides by IMAC can be employed.[224,233] The technique involves chelation of metals such as Fe^{3+} or Ga^{3+} onto a chromatographic support consisting of iminodiacetic acid or nitrilotriacetic acid.[236-238] Phosphopeptides, being acidic by virtue of the phosphate group, bind with some selectively over nonphosphopeptides. Fractions enriched for phosphopeptides are then eluted by phosphate or increased pH. While the method is somewhat selective for phosphopeptides, other peptides, particularly those containing strings of acidic amino acids or histidine, are also enriched. The method has been applied on-line to MS in an integrated peptide enrichment/separation system consisting of a tandem IMAC/RP column configuration.[215,224,235]

d. Reversed-Phase HPLC. Reversed-phase HPLC fractionation of phosphopeptides is reproducible, simple, and does not require specialized equipment.[225,231,239] In RP-HPLC, ^{32}P-labeled phosphopeptides are separated on the basis of their hydrophobicity and fractions collected for Cerenkov counting (i.e., without scintillation fluid so that the samples may be further analyzed by MS). A graph of Cerenkov counts versus elution time reveals the number of radioactive fractions that can then be analyzed by the MS methods described below.[231,239] A disadvantage of RP-HPLC over 2DPP mapping is that very hydrophilic phosphopeptides may not stick to the column and thus will elute in the column flow-through. Conversely, very hydrophobic peptides may not elute until the end of a gradient and will be obscured by the polymeric contaminants that often elute at high acetonitrile concentrations or they may not elute at all. It is therefore possible that some of the phosphopeptides in a sample will go undetected. Generally, the resolution of RP-HPLC is also inferior to the resolution achieved by the two-dimensional peptide mapping technique. An additional note of caution is that phosphopeptides will stick to metal surfaces. Significant sample losses can occur if standard metal injectors are used. Even considering these disadvantages, RP-HPLC for phosphopeptide analysis is popular because of the ease with which RP-HPLC systems are connected on-line to ESI mass spectrometers. The use of a mass spectrometer connected on-line to the HPLC system makes it possible to detect and characterize phosphopeptides in sample mixtures, even if the analyte is not radiolabeled. This

is achieved by implementing one of several possible phosphate-specific diagnostic ion scans, which include precursor ion scans,[240] neutral loss scans,[241] and in-source dissociation.[242]

e. Capillary Electrophoresis. Recently, two methods using capillary electrophoresis (CE) for analysis of phosphopeptides via ESI were published.[141,224] The method by Figeys et al. incorporated a solid-phase extraction (SPE) capillary zone electrophoresis (CZE) device for peptide concentration/separation on-line with ESI-MS and an algorithm written in Instrument Control Language (ICL) that modulated the electrophoretic conditions in a data-dependent manner to optimize available time for the generation of high-quality CID spectra of peptides in complex samples. The data-dependent modulation of the electric field significantly expanded the analytical window for each peptide analyzed and enhanced the sensitivity by reducing the CE voltage and thus the flow into the ESI source. The technique was applied to the analysis of in vivo phosphorylation sites of endothelial nitric oxide synthase (eNOS) demonstrating the power of the method for the MS/MS analysis of minor peptide species in complex samples such as phosphopeptides generated by the proteolytic digestion of a large protein, eNOS, phosphorylated at low stoichiometry. The second application of CE for analysis of phosphopeptides took a slightly different approach technically but also was concerned with analysis of minor phosphopeptide components produced by proteolysis of large proteins.[243] The method used Fe(III) immobilized metal-ion affinity chromatography (IMAC)-CE-electrospray ionization MS to analyze subpicomole analysis of phosphopeptides. The IMAC resin was packed directly at the head of the CE column, and after the phosphopeptides were bound to the resin and washed, they were eluted at high pH and separated by CE. Advantages of this approach include (i) selective retention and preconcentration of phosphopeptides; (ii) a prewash of the sample to remove salts and buffers that are not suited for CE separation or ESI operation; (iii) ease of construction; and (iv) adaptation to commercial CE instruments without any modifications.

4. Phosphopeptide Sequence Determination

There are number of different mass spectrometric methods for determining which amino acid residue(s) in a peptide are phosphorylated, and they fall into two general themes. The first relies on the chemical lability of the phosphoester bonds in phosphoserine, -threonine, and -tyrosine. These phosphoester bonds can easily be induced to fragment in a collision cell or the ion source of an ESI instrument or during PSD in a MALDI-MS, resulting in loss of phosphate from the peptide. Phosphopeptides that lose phosphate due to any of these processes can then be identified by any of several phosphate-specific diagnostic ion scans. The second theme relies on the detection of the mass added to a peptide by the phosphate group. Typically, in protein phosphorylation studies, the amino acid sequence of the protein investigated is known. Therefore, phosphopeptides derived from the protein can, in principle, be detected by a net mass differential of 80 u that occurs when phosphate is added to serine, threonine, or tyrosine. Thus, a peptide mass map of the proteolytically fragmented phosphoprotein can potentially identify the phosphorylated peptide by comparison to the known protein sequence. Neither method however identifies the phosphorylated amino acid residue(s) within the peptide directly. In cases where the peptide sequence contains only a single possible phosphorylation site, the phosphorylated residue is effectively located by default. If this is not the case, then tandem MS is necessary to locate the phosphorylated amino acid residue. In general, methods that produce some sort of phosphate-specific ion (i.e., a diagnostic ion) are useful or even essential for the detection of phosphopeptides in mixtures in cases in which incorporation of ^{32}P is not possible or in which the radiolabel has decayed past the point of detection.[244] However, such scans can also be used on radiolabeled phosphopeptides because the contribution to the mass of the phosphopeptide from the radioactive isotope of phosphate is so small that it can be ignored. If researchers are concerned about contaminating a mass spectrometer with ^{32}P samples, they can easily avoid contamination by waiting a sufficient number of half-lives (2 weeks for ^{32}P) before conducting mass spectrometric experiments. Several types of MS-based approaches to phosphopeptide analysis have been developed and applied. The most commonly used ones are described in the following.

a. In-Source CID. If phosphopeptide ions are fragmented in negative ion mode, $H_2PO_4^-$ (97 u), PO_3^- (79 u), and PO_2^- (63 u) are detected as phosphate-specific diagnostic ions. Under low-energy CID conditions, phosphotyrosine will be observed to generate the last two of these three ions, PO_3^- (79 u) and PO_2^- (63 u) but not $H_2PO_4^-$ (97 u). These phosphate-specific diagnostic ions can be selectively monitored to identify phosphopeptides.[245,246] When in-source CID is combined on-line with HPLC, a chromatographic trace is established that identifies the elution time of a phosphopeptide. Carr and coworkers developed a negative ion LC-MS protocol that monitors the phosphopeptide-specific reporter ions and determines the phosphopeptide molecular weight in the same scan.[245] This was accomplished by use of a high orifice potential across the two skimmers prior to Q1 in a triple quadrupole instrument while the low *m/z* range is scanned for the diagnostic ions. The orifice potential is then lowered to a voltage that does not induce fragmentation and the high *m/z* range is scanned. A similar experiment can be done on instruments where a heated capillary replaces the first skimmer.[246,247] This method, an extension of that of Hunter and Games,[246] uses an alternating scan approach where selected ion monitoring of appropriate diagnostic ions at a high octapole offset voltage is followed by two full scans. The first full scan is conducted at the same high offset voltage as the SIM experiment providing signals for the deprotonated phosphopeptide molecular ion and the phosphopeptide molecular ion minus phosphate. Finally a second full scan is done at a normal octapole offset voltage to provide a reference to the full scan at high octapole offset. This series of three MS scans

is repeated continuously throughout the LC separation. Such an experiment provides the same information as the methods of Carr and co-workers,[245] but because selected ion monitoring (SIM) is used rather than scanning to detect diagnostic ions, one cycle of scans is faster and potentially more sensitive. The first full scan at high octapole offset is compared to the full scan at low octapole offset to provide a clue as to which peptide ion is phosphorylated. This is useful to distinguish phosphopeptides from nonphosphorylated peptides in case they coelute from the microcapillary columns. Such techniques are generally capable of identifying low femtomole amounts of standard applied to a capillary column. Frequently, the detection sensitivity with "real" in vitro or in vivo phosphorylated samples is lower and can drop into the picomole range.

Of course it would be advantageous to also determine the amino acid sequence of the detected phosphopeptide and the phosphorylated residue(s) in the same negative ion LC-MS experiment. Unfortunately, this has been difficult to achieve because negative ion CID spectra generally produce insufficient fragment ions for sequence elucidation. Attempts to switch between negative ion mode for phosphopeptide detection and positive ion mode for peptide product ion scanning in the same LC-MS experiment have been technically difficult in scanning mass spectrometers (such as quadrupole instruments) due to the time required to switch between positive and negative ion mode in real time. It appears that such experiments might be easier to carry out in nonscanning mass spectrometers.

b. Neutral Loss Scanning. Neutral loss scanning carried out in positive ion mode with ESI in a triple quadrupole MS for phosphopeptide detection and analysis was first described by Covey et al.[241] and further developed by Huddelston et al.[245] Instead of using Q1 to select specific ions for fragmentation in Q2, Q1 is scanned coordinately with Q3 but with an offset in *m/z* value. The offset in the *m/z* values measured by the two quadrupoles at any given time is constant and corresponds to the loss of phosphate by CID in Q2. This loss reflects the facile loss of phosphate from phosphoserine and phosphothreonine due to a process known as β-elimination. Only phosphothreonine and phosphoserine (not shown) may undergo neutral loss of H_3PO_4 (98) by β-elimination as shown in Figure 7 but not phosphotyrosine.[248] This is because the α-carbon proton that is abstracted by a lone pair of electrons from the phosphate moiety is too far removed for facile loss (Figure 7B); i.e., in tyrosine phosphate the phosphate group is no longer in the β position relative to the α-carbon proton. Therefore, neutral loss scanning for a loss of phosphate from a $[M + 2H]^{2+}$ phosphopeptide ion requires an offset value is 49 *m/z*. The method has not been as popular as the aforementioned in-source CID methods because of the propensity of false-positive signals and the need to know the charge state of the ion losing phosphate. It is an advantage of the method that it is carried out in positive ion mode and can therefore be used with data-dependent scanning to acquire CID spectra in the same experiment using

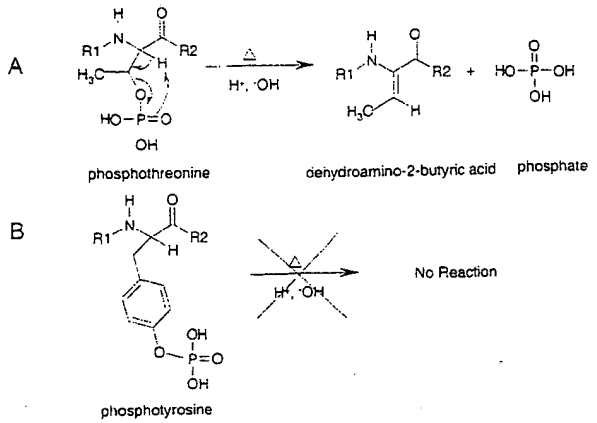

Figure 7. Mechanism for β-elimination of phosphate. Illustration of chemical mechanism for (A) β-elimination of phosphate from phosphothreonine to produce dehydroamino-2-butyric acid (scheme is also valid for phosphoserine where dehydroalanine forms) and phosphate and (B) absence of the same mechanistic pathway for β-elimination of phosphate from phosphotyrosine.

the detection of a neutral loss of phosphate as a trigger to initiate CID.

c. Precursor Ion Scanning. In this method, negative ion ESI is carried out with continuous scanning of Q1 and Q3 maintained at a fixed *m/z* ratio. Ions are fragmented in Q2, and Q3 passes only one ion; in the case of phosphopeptide analysis this is usually 79 *m/z* (i.e., loss of PO_3^-). Consequently, the resultant mass spectrum shows only ions that lost 79 *m/z*.[140,249,250] This greatly simplifies mixture analysis and is best done during direct infusion with a nanospray source. Again as described for the phosphate diagnostic ion scans, there is a problem associated with sequencing in positive ion mode immediately after detecting the loss of phosphate in negative ion mode.

d. Product Ion Scanning. Often the information obtained by the specific scanning methods described above is not sufficient for identification of the phosphorylated residue in a phosphopeptide. In fact, the above methods of in-source CID, neutral loss, and precursor ion scanning are designed to distinguish phosphopeptides from nonphosphopeptides and to potentially indicate the phosphopeptide mass rather than to provide sequence information. Consequently, these methods can only successfully identify a phosphorylated residue if the peptide sequence is known and contains only one copy of one of the hydroxyl amino. If there is more than one possible amino acid residue present in the peptide that can be phosphorylated, then it is necessary to acquire tandem mass spectra for either manual or algorithm-based sequence interpretation.[231,239]

As a general trend for low-energy CID of phosphopeptides, it has been observed that phosphate tends to be lost from phosphoserine more readily than phosphothreonine and from phosphothreonine more readily than from phosphotyrosine. Phosphate is generally eliminated from shorter phosphopeptides more readily than from longer phosphopeptides because roughly the same amount of energy for collision is dispersed across fewer bonds. Interestingly, it is

rare to observe the immonium ions for phosphoamino acids that form as a result of dehydroalanine and dehydroamino-2-butyric acid breaking down after loss of phosphate. However, using an ion trap mass spectrometer and monitoring the CID of a phosphopeptide ion, dehydroamno-2-butyric acid (Figure 7A) was observed in place of threonine in the peptide fragment ion.[250]

e. Post-Source Decay. Meta-stable decay of phosphopeptides has been observed during PSD-MALDI-TOF and provides a method to sequence peptides in a single-stage instrument. While not popular for reasons cited above, the method has been successfully applied to the analysis of phosphopeptides.[236,251]

f. Enzymatic Dephosphorylation. Phosphatases can be used to identify phosphopeptides in a mixture of predominantly nonphosphopeptides. Typically, as a first step, the peptide masses resulting from proteolytic digestion of the phosphoprotein are acquired in a MALDI-TOF instrument. Second, the same sample is treated with phosphatase to remove phosphate selectively from the phosphopeptide(s) and the masses acquired again. Any peptide mass that decreases by 80 u as a result of the phosphatase treatment will be designated a potential phosphopeptide. An advantage to conducting such an experiment by MALDI-MS is that peptide ions produced tend to be singly charged rather than multiply protonated and that the phosphatase reaction can be carried out directly on the MALDI probe.[252,253] A similar approach has also been developed for ESI-MS.[254] In this method, a enzyme microreactor consisting of an immobilized tyrosine phosphatase was used to dephosphorylate peptides on-line prior to analysis by CE-MS/MS. As for the MALDI-based method, phosphopeptides in the peptide mixture were characterized by a mass difference of 80 u when the MS data obtained with and without enzyme reactor were compared. The method has the additional advantage that the peptides were also undergoing a mobility shift in CE upon dephosphorylation, further confirming their identity as phosphopeptides and that the phosphorylated or dephosphorylated species of the phosphopeptide could be directly subjected to CID for further characterization and location of the phosphorylated residue, respectively.

IV. Conclusions

Traditionally, advances in mass spectrometric methods for the analysis of proteins and peptides were driven primarily by the need to identify and analyze purified proteins faster, more sensitively, and more reliably. The advent of complete genome sequences has accelerated incremental improvements in mass spectrometric methods for protein identification and analysis and also catalyzed a new research method.

Incremental improvements have been accelerated because the genome sequence databases contain the sequence information for every protein potentially expressed by that genome. Consequently, proteins isolated from species with complete sequence databases are no longer identified by de novo sequencing but rather by correlating idiotypic information extracted from the intact polypeptide or a peptide fragment thereof with the sequence database. Currently, MS is the method of choice for the generation of data for sequence database searching and therefore a cornerstone of analytical protein chemistry.

The genomics revolution has also catalyzed a new research method we have termed discovery science.[4] Discovery science enumerates the elements of a biological system irrespective of any hypotheses of how the system functions. Discovery science complements the traditional hypothesis driven method to biological research, and proteomics is an essential component of discovery science. The initial efforts of proteomics have been focused on the identification of the proteins expressed by a cell or tissue a process that can be described as descriptive proteomics. More recently, the focus has shifted to the development of methods capable of measuring, on a proteome-wide scale, properties of proteins that reflect the function and dynamics of proteins. These include the quantity, the state of modification, the specific activity, and the association of a protein with other macromolecules. Many of these methods depend on MS and are currently being rapidly further developed. The potential for method refinement, for developing methods to uncover new types of information, and for the power of the current methods to dissect biological systems at a molecular level make MS and proteomics among the most exciting, dynamic, and important research themes at the present time.

Acknowledgments. D.R.G. thanks Dr. J. D. Watts for critical opinions and edification in protein phosphorylation analysis over the last 3 yr. More information about the author's research can be found at http://www.systemsbiology.org. For issues concerning MS, the American Society for Mass Spectrometry has an excellent website at http://www.asms.org.

V. References

(1) Rowen, L.; Mahairas, G.; Hood, L. *Science* **1997**, *278*, 605.
(2) Fraser, C. M.; Fleischmann, R. D. *Electrophoresis* **1997**, *18*, 1207.
(3) Harry, J. L.; Wilkins, M. R.; Herbert, B. R.; Packer, N. H.; Gooley, A. A.; Williams, K. L. *Electrophoresis* **2000**, *21*, 1071.
(4) Aebersold, R.; Hood, L. E.; Watts, J. D. *Nat. Biotechnol.* **2000**, *18*, 359.
(5) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837.
(6) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. *Science* **1996**, *274*, 563.
(7) Consortium, *C. Elegans. Sci.* **1998**, *282*, 2012.
(8) Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; George, R. A.; Lewis, S. E.; Richards, S.; Ashburner, M.; Henderson, S. N.; Sutton, G. G.; Wortman, J. R.; Yandell, M. D.; Zhang, Q.; Chen, L. X.; Brandon, R. C.; Rogers, Y. H.; Blazej, R. G.; Champe, M.; Pfeiffer, B. D.; Wan, K. H.; Doyle, C.; Baxter, E. G.; Helt, G.; Nelson, C. R.; Gabor-Miklos, G. L.; Abril, J. F.; Agbayani, A.; An, H. J.; Andrews-Pfannkoch, C.; Baldwin, D.; Ballew, R. M.; Basu, A.; Baxendale, J.; Bayraktaroglu, L.; Beasley, E. M.; Beeson, K. Y.; Benos, P. V.; Berman, B. P.; Bhandari, D.; Bolshakov, S.; Borkova, D.; Botchan, M. R.; Bouck, J.; Brokstein, P.; Brottier, P.; Burtis, K. C.; Busam, D. A.; Butler, H.; Cadieu, E.; Center, A.; Chandra, I.; Cherry, J. M.; Cawley, S.; Dahlke, C.; Davenport, L. B.; Davies, P.; de Pablos, B.; Delcher, A.; Deng, Z.; Mays, A. D.; Dew, I.; Dietz, S. M.; Dodson, K.; Doup, L. E.; Downes, M.; Dugan-Rocha, S.; Dunkov, B. C.; Dunn, P.; Durbin, K. J.; Evangelista, C. C.; Ferraz, C.; Ferriera, S.; Fleischmann, W.; Fosler, C.; Gabrielian, A. E.; Garg, N. S.; Gelbart, W. M.; Glasser, K.; Glodek, A.; Gong, F.; Gorrell, J. H.; Gu, Z.; Guan, P.; Harris, M.; Harris, N. L.; Harvey, D.; Heiman, T. J.; Hernandez, J. R.; Houck, J.; Hostin, D.; Houston, K. A.; Howland, T. J.; Wei, M.

H.; Ibegwam, C.; Jalali, M.; Kalush, F.; Karpen, G. H.; Ke, Z.; Kennison, J. A.; Ketchum, K. A.; Kimmel, B. E.; Kodira, C. D.; Kraft, C.; Kravitz, S.; Kulp, D.; Lai, Z.; Lasko, P.; Lei, Y.; Levitsky, A. A.; Li, J.; Li, Z.; Liang, Y.; Lin, X.; Liu, X.; Mattei, B.; McIntosh, T. C.; McLeod, M. P.; McPherson, D.; Merkulov, G.; Milshina, N. V.; Mobarry, C.; Morris, J.; Moshrefi, A.; Mount, S. M.; Moy, M.; Murphy, B.; Murphy, L.; Muzny, D. M.; Nelson, D. L.; Nelson, D. R.; Nelson, K. A.; Nixon, K.; Nusskern, D. R.; Pacleb, J. M.; Palazzolo, M.; Pittman, G. S.; Pan, S.; Pollard, J.; Puri, V.; Reese, M. G.; Reinert, K.; Remington, K.; Saunders, R. D.; Scheeler, F.; Shen, H.; Shue, B. C.; Siden-Kiamos, I.; Simpson, M.; Skupski, M. P.; Smith, T.; Spier, E.; Spradling, A. C.; Stapleton, M.; Strong, R.; Sun, E.; Svirskas, R.; Tector, C.; Turner, R.; Venter, E.; Wang, A. H.; Wang, X.; Wang, Z. Y.; Wassarman, D. A.; Weinstock, G. M.; Weissenbach, J.; Williams, S. M.; Woodage, T.; Worley, K. C.; Wu, D.; Yang, S.; Yao, Q. A.; Ye, J.; Yeh, R. F.; Zaveri, J. S.; Zhan, M.; Zhang, G.; Zhao, Q.; Zheng, L.; Zheng, X. H.; Zhong, F. N.; Zhong, W.; Zhou, X.; Zhu, S.; Zhu, X.; Smith, H. O.; Gibbs, R. A.; Myers, E. W.; Rubin, G. M.; Venter, J. C. *Science* **2000**, *287*, 2185.
(9) Butler, D.; Pockley, P. *Nature* **2000**, *404*, 534.
(10) Pennisi, E. *Science* **2000**, *288*, 239.
(11) Pennisi, E. *Science* **2000**, *289*, 2304.
(12) National Cancer Institute. dbEST, 2000, www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.
(13) Hofmann, K.; Bucher, P.; Falquet, L.; Bairoch, A. *Nucleic Acids Res.* **1999**, *27*, 215.
(14) Henikoff, S.; Henikoff, J. G.; Pietrokovski, S. *Bioinformatics* **1999**, *15*, 471.
(15) Skolnick, J.; Fetrow, J. S.; Kolinski, A. *Nat. Biotechnol.* **2000**, *18*, 283.
(16) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. *Science* **1999**, *285*, 751.
(17) Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. *Nature* **1999**, *402*, 86.
(18) Wilkins, M. R.; Sanchez, J. C.; Gooley, A. A.; Appel, R. D.; Humphery-Smith, I.; Hochstrasser, D. F.; Williams, K. L. *Biotechnol. Genet. Eng. Rev.* **1996**, *13*, 19.
(19) Wasinger, V. C.; Cordwell, S. J.; Cerpa-Potjak, A.; Yan, J. X.; Gooley, A. A.; Wilkins, M. R., Duncan, M. W.; Harris, R. Williams, K. L.; Humphrey-Smith, I. *Electrophoresis* **1995**, *16*, 1090.
(20) Hochstrasser, D. F. *Clin. Chem. Lab. Med.* **1998**, *36*, 825.
(21) Loo, R. R. O.; Stevenson, T. I.; Mitchell, C.; Loo, J. A.; Andrews, P. C. *Anal. Chem.* **1996**, *8*, 1910.
(22) Haynes, P. A.; Gygi, S. P.; Figeys, D.; Aebersold, R. *Electrophoresis* **1998**, *19*, 1862.
(23) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64.
(24) Cole, R. B. *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation and Applications*; Wiley: New York, 1997.
(25) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1995**, *60*, 2299.
(26) Barber, M.; Bordoli, R. S.; Sedgwick, R. D.; Tyler, A. N. J. *Chem. Soc. Commun.* **1981**, 325.
(27) Jardine, I. *Methods Enzymol.* **1990**, *193*, 441.
(28) Chaurand, P.; Luetzenkirchen, F.; Spengler, B. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 91.
(29) Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* **1993**, *65*, 425–38.
(30) Chait, B. T.; Kent, S. B. *Science* **1992**, *257*, 1885.
(31) Loo, J. A. *Bioconjugate Chem.* **1995**, *6*, 644.
(32) Goodlett, D. R.; Ogorzalek-Loo, R. R.; Loo, J. A.; Wahl, J. H.; Udseth, H. R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 614.
(33) Ganem, B.; Li, Y.-T.; Henion, J. D. *J. Am. Chem. Soc.* **1991**, *113*, 7818.
(34) Loo, J. A. *Mass Spectrom. Rev.* **1997**, *16*, 1.
(35) Wood, T. D.; Chorush, R. A.; Wampler, F. M., III; Little, D. P.; O'Connor, P. B.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2451.
(36) Anderegg, R. J.; Wagner, D. S.; Stevenson, C. L.; Borchardt, R. T. *J. Am. Soc. Mass Spectrom.* **1994**, *4*, 425.
(37) Akashi, S.; Naito, Y.; Takio, K. *Anal. Chem.* **1999**, *71*, 4974.
(38) Loo, J. A.; Loo, R. R.; Udseth, H. R.; Edmonds, C. G.; Smith, R. D. *Rapid Commun. Mass Spectrom.* **1991**, *5*, 101.
(39) Katta, V.; Chait, B. T. *Rapid Commun. Mass Spectrom.* **1991**, *5*, 214.
(40) Mirza, U. A.; Cohen, S. L.; Chait, B. T. *Anal. Chem.* **1993**, *65*, 1.
(41) Konermann, L.; Collings, B. A.; Douglas, D. J. *Biochemistry* **1997**, *36*, 5554.
(42) Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5802.
(43) Cohen, S. L.; Padovan, J. C.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 574.
(44) Rappsilber, J.; Siniossoglou, S.; Hurt, E. C.; Mann, M. *Anal. Chem.* **2000**, *72*, 267.
(45) Kennedy, R.; J. W. Jorgenson, J. W. *Anal. Chem.* **1991**, *63*, 1467.
(46) Hunt, D. F.; Alexander, J. E.; McCormack, A. L.; Martino, P. A.; Michel, H.; Shabanowitz, J.; Sherman, N.; Moseley, M. A.; Jorgenson, J. W.; Tomer, K. B. *Techniques in Protein Chemistry II*; Academic Press: New York, 1991; p 441.
(47) Lee, N.; Goodlett, D. R.; Marquardt, H.; Geraghty, D. E. *J. Immunol.* **1998**, *160*, 4951.
(48) Mosely, M. A.; Deterding, L. J.; Tomer, K. B.; Jorgenson, J. W. *Anal. Chem.* **1991**, *63*, 1467.
(49) Wahl, J. H.; Goodlett, D. R.; Udseth, H. R.; Smith, R. D. *Anal. Chem.* **1992**, *64*, 3194.
(50) Wahl, J. H.; Goodlett, D. R.; Udseth, H. R.; Smith, H. R. *Electrophoresis* **1993**, *14*, 448.
(51) Wahl, J. H.; Gale, D. C.; Smith, R. D. *J. Chromatogr. A* **1994**, *659*, 217.
(52) Wilm, M. S.; Mann, M. *Int. J. Mass Spectrom. Ion Processes* **1994**, *136*, 167.
(53) Wilm, M.; Mann, M. *Anal. Chem.* **1996**, *68*, 1.
(54) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233.
(55) Matsudaira, P. *J. Biol. Chem.* **1987**, *262*, 10035.
(56) Lazar, I. M.; Ramsey, R. S.; Sundberg, S.; Ramsey, J. M. *Anal. Chem.* **1999**, *71*, 3627.
(57) Carr, S. A.; Annan, R. S. *Current Protocols in Molecular Biology*; John Wiley & Sons: New York, 1997; p 10.21.1.
(58) Davis, M. T.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 1059.
(59) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69*, 767.
(60) Figeys, D.; Aebersold, R. *Electrophoresis* **1998**, *19*, 885.
(61) Figeys, D.; Gygi, S. P.; McKinnon, G.; Aebersold, R. *Anal. Chem.* **1998**, *70*, 3728.
(62) Belov, M. E.; Gorshkov, M. V.; Udseth, H. R.; Anderson, G. A.; Tolmachev, A. V.; Prior D. C.; Harkewicz, R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 19.
(63) Zhang, B.; Foret, F.; Karger, B. L. *Anal.Chem.* **2000**, *72*, 1015.
(64) Yost, R. A.; Enke, C. G. *Anal. Chem.* **1979**, *51*, 1251A.
(65) Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Science* **1990**, *248*, 201.
(66) Schwartz, J. C.; Jardine, I. *Methods Enzymol.* **1996**, *270*, 552.
(67) Jonscher, K. R.; Yates, J. R., III. *Anal. Biochem.* **1997**, *244*, 1.
(68) Morris, H. R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R. S.; Hoyes, J.; Bateman, R. H. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 889.
(69) Borchers, C.; Parker, C. E.; Deterding, L. J.; Tomer, K. B. *J. Chromatogr. A* **1999**, *854*, 119.
(70) Borchers, C.; Peter, J. F.; Hall, M. C.; Kunkel, T. A.; Tomer, K. B. *Anal. Chem.* **2000**, *72*, 1163.
(71) Fitzgerald, M. C.; Chernushevich, I. I.; Standing, K. G.; Whitman, C. P.; Kent, S. B. *Proc. Natl. Acad. Sci U.S.A.* **1996**, *93*, 6851.
(72) Tanaka, K.; Kawatoh, E.; Ding, L.; Smith, A. J.; Kumashiro, S. *Proceedings of 47th American Society for Mass Spectrometry and Allied Topics*; 1999; TP086.
(73) Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. *Anal. Chem.* **2000**, *72*, 552.
(74) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1.
(75) Wood, T. D.; Guan, Z.; Borders, C. L.; Chen, L. H.; Kenton, G. L.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 3362.
(76) Qin, J.; Chait, B. T. *Anal. Chem.* **1997**, *69*, 4002.
(77) Qin, J.; Fenyo, D.; Zhao, Y.; Hall, W. W.; Chao, D. M.; Wilson, C. J.; Young, R. A.; Chait, B. T. *Anal. Chem.* **1997**, *69*, 3995.
(78) Conuterman, A. E.; Valentine, S. J.; Srebalus, C. A.; Henderson, S. C.; Hoagland, C. S.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 743.
(79) Burlingame, A. L.; Boyd, R. K.; Gaskell, S. J. *Anal. Chem.* **1996**, *68*, 599R.
(80) Burlingame, A. L.; Boyd, R. K.; Gaskell, S. J. *Anal. Chem.* **1998**, *68*, 647R.
(81) Hewick, R. M.; Hunkapiller, M. W.; Hood, L. E.; Dreyer, W. J. *J. Biol. Chem.* **1981**, *256*, 7990.
(82) Aebersold, R. H.; Leavitt, J.; Saavedra, R. A.; Hood, L. E.; Kent, S. B. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6970.
(83) Larive, C. K.; Lunte, S. M.; Zhong, M.; Perkins, M. D.; Wilson, G. S.; Gokulrangan, G.; Williams, T.; Afroz, F.; Schoneich, C.; Derrick, T. S.; Middaugh, C. R.; Bogdanowich-Knipp, S. *Anal. Chem.* **1999**, *71*, 398R.
(84) Lamond, A. I.; Mann, M. *Trends Cell Biol.* **1997**, *7*, 139.
(85) Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791.
(86) Moritz, R. L.; Eddes, J.; Ji, H.; Reid, G. E.; Simpson, R. J. *Techniques in Protein Chemistry VI*; Academic Press: San Diego, 1995; p 311.
(87) Aebersold, R.; Leavitt, J. *Electrophoresis* **1990**, *11*, 517.
(88) Aebersold, R. H.; Teplow, D. B.; Hood, L. E.; Kent, S. B. H. *J. Biol. Chem.* **1986**, *261*, 4229.
(89) Tonella, L.; Walsh, B. J.; Sanchez, J. C.; Ou, K.; Wilkins, M. R.; Tyler, M.; Frutiger, S.; Gooley, A. A.; Pescaru, I.; Appel, R. D.;

Yan, J. X.; Bairoch, A.; Hoogland, C.; Morch, F. S.; Hughes, G. J.; Williams, K. L.; Hochstrasser, D. F. *Electrophoresis* **1998**, *19*, 1960.
(90) Joubert, R.; Brignon, P.; Lehmann, C.; Monribot, C.; Gendre, F.; Boucherie, H. *Yeast* **2000**, *16*, 511.
(91) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci U.S.A.* **1996**, *93*, 14440.
(92) Langen, H.; Takacs, B.; Evers, S.; Berndt, P.; Lahm, H. W.; Wipf, B.; Gray, C.; Fountoulakis, M. *Electrophoresis* **2000**, *21*, 411.
(93) Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J. C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphery-Smith, I.; Williams, K. L.; Hochstrasser, D. F. *Biotechnology* **1996**, *14*, 61.
(94) Wilkins, M. R.; Gasteiger, E.; Wheeler, C. H.; Lindskog, I.; Sanchez, J. C.; Bairoch, A.; Appel, R. D.; Dunn, M. J.; Hochstrasser, D. F. *Electrophoresis* **1998**, *19*, 3199.
(95) Wilkins, M. R.; Yan, J. X.; Gooley, A. A. *Methods Mol. Biol.* **1999**, *112*, 445.
(96) Gobom, J.; Nordhoff, E.; Mirgorodskaya, E.; Ekman, R.; Roepstorff, P. *Anal. Chem.* **1999**, *71*, 919.
(97) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C. Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011.
(98) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58.
(99) Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397.
(100) Pappin, D. J. *Methods Mol. Biol.* **1997**, *64*. 165.
(101) Jensen, O. N.; Larsen, M. R.; Roepstorff, P. *Proteins Suppl.* **1998**, *2*, 74.
(102) Jensen, O. N.; Podtelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997**, *69*, 4741.
(103) Eriksson, J. Chait, B. T.; Fenyo, D. *Anal. Chem.* **2000**, *72*, 999.
(104) Cordwell, S. J.; Wilkins, M. R.; Cerpa-Poljak, A.; Gooley, A. A.; Duncan, M.; Williams, K. L.; Humphery-Smith, I. *Electrophoresis* **1995**, *16*, 438.
(105) Fenyo, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998.
(106) Jensen, O. N.; Podtelejnikov, A.; Mann, M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1371.
(107) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871.
(108) Takach, E. J.; Hines, W. M.; Patterson, D. H.; Juhasz, P.; Falick, A. M.; Vestal, M. L.; Martin, S. A. *J. Protein Chem.* **1997**, *16*, 363.
(109) Bruce, J. E.; Anderson, G. A.; Wen J.; Harkewitz, R.; Smith, R. D. *Anal. Chem.* **1999**, *71*, 2595.
(110) Goodlett, D. R.; Bruce, J. E.; Anderson, G. A.; Rist, B.; Pasa-Tolic, L.; Fiehn, O.; Smith, R. D.; Aebersold, R. *Anal. Chem.* **2000**, *72*, 1112.
(111) Pappin, D. J. C.; Rahman, D.; Hansen, H. F.; Bartlet-Jones, M.; Jeffery, W.; Bleasby, A. J. *Mass Spectrometry in the Biological Sciences*; Humana Press: Totowa, 1995; p 135.
(112) Craig, A. G.; Fischer, W. H.; Rivier, J. E.; McIntosh, J. M.; Gray, W. R. *Techniques in Protein Chemistry VI*; Academic Press: San Diego, 1990; p 31.
(113) Sechi, S.; Chait, B. T. *Anal. Chem.* **1998**, *70*, 5150.
(114) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Protein Sci.* **1994**, *3*, 1347.
(115) Jensen, O. N.; Vorm, O.; Mann, M. *Electrophoresis* **1996**, *17*, 938.
(116) Woods, A. S.; Huang, A. Y. C.; Cotter, R. J.; Pasternack, G. R.; Pardoll, D. M.; Jaffee, E. M. *Anal. Biochem.* **1995**, *226*, 15.
(117) Patterson, S. D. *Electrophoresis* **1995**, *16*, 1104.
(118) Wiley: W. C.; McLaren, I. H. *Rev. Sci. Instrum.* **1953**, *26*, 1150.
(119) Vestal, M. L.; Juhasz, P.; Martin, S. A. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1044.
(120) Brown, R. S.; Lennon, J. J. *Anal. Chem.* **1995**, *67*, 1998.
(121) Urquhart, B. L.; Atsalos, T. E.; Roach, D.; Basseal, D. J.; Bjellqvist, B.; Britton, W. L.; Humphery-Smith, I. *Electrophoresis* **1997**, *18*, 1384.
(122) Zubarev, R. A.; Hakansson, P.; Sundqvist, B. *Anal. Chem.* **1996**, *68*, 4060.
(123) Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* **1991**, *63*, 2488.
(124) Light-Wahl, K. J.; Loo, J. A.; Edmonds, C. G.; Smith, R. D.; Witkowska, H. E.; Shackleton, C. H.; Wu, C. S. *Biol. Mass Spectrom.* **1993**, *22*, 112.
(125) Lennon, J. J.; Walsh KA *Protein Sci.* **1997**, *6*, 2446.
(126) Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. *Anal. Chem.* **2000**, *72*, 563.
(127) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676.
(128) Suckau, D.; Cornett, D. S. *Analusis* **1998**, *26*, M18.
(129) Biemann, K. *Methods Enzymol.* **1990**, *193*, 886.
(130) Roepstorff, P.; Fohlman, J. *J. Biomed. Mass Spectrom.* **1984**, *11*, 601.
(131) Harrison, A. G.; Csizmadia, I. G.; Tang, T. H. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 427.
(132) Loo, J. A.; Edmonds, C. G. Smith, R. D. *Anal. Chem.* **1993**, *65*, 425.
(133) Schwartz, B. L.; Bursey, M. M. *Biol. Mass Spectrom.* **1992**, *21*, 92.
(134) Goodlett, D. R.; Gale, D. C.; Guiles, S.; Crowther, J. *Encyclopedia of Analytical Chemistry*; John Wiley: New York, 2000.
(135) Mak, M.; Mezo, G.; Skribanek, Z.; Hudecz, F. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 837.
(136) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365.
(137) Spengler, B.; Luetzenkirchen, F.; Metzger, S.; Chaurand, P.; Kaufmann, R.; Jeffery, W.; Bartlet-Jones, M.; Pappin, D. J. C. *Int. J. Mass Spectrom. Ion Processes* **1997**, *169/170*, 127.
(138) Griffiths, W. J. *EXS* **2000**, *88*, 69.
(139) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466.
(140) Wilm, M.; Neubauer, G.; Mann, M. *Anal. Chem.* **1996**, *68*, 527.
(141) Figeys, D.; Corthals, G. L.; Gallis, B.; Goodlett, D. R.; Ducret, A.; Corson, M. A.; Aebersold, R. *Anal. Chem.* **1999**, *71*, 2279.
(142) Stahl, D. C.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 532.
(143) Shabanowitz, J.; Settlage, R. E.; Marto, J. A.; Christian, R. E.; White, F. M.; Russo, P. S. W.; Martin, S. E.; Hunt, D. F. *Mass Spectrometry in Biology and Medicine*; Human Press: Totowa, 2000; p 163.
(144) Ducret, A.; van Oostveen, I.; Eng, J. K.; Yates, J. R., III; Aebersold, R. *Protein Sci.* **1998**, *7*, 706.
(145) Courchesne, P. L.; Jones, M. D.; Robinson, J. R.; Spahr, C. S.; McCracken, S.; Bentley, D. L.; Luethy, R.; Patterson, S. D. *Electrophoresis* **1998**, *19*, 956.
(146) Goodlett, D. R.; Wahl, J. H.; Udseth, H. R.; Smith, R. D. *J. Microcolumn Sep.* **1993**, *5*, 57.
(147) Susin, S. A.; Lorenzo, H. K.; Zamzmi, N.; Marzo, I.; Snow, B. E.; Brothers, G. M.; Mangion, J.; Jacotot, E.; Costantini, P.; Loeffler, M. M.; Larochette, N.; Goodlett, D. R.; Aebersold, R.; Siderovski, D. P.; Penninger, J. M.; Kroemer, G. *Nature* **1999**, *397*, 441.
(148) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69*, 767.
(149) Chait, B. T.; Wang, R.; Beavis, R. C.; Kent, S. B. H. *Science* **1993**, *262*, 89.
(150) Wang, R.; Chait, B. T.; Kent, S. B. H. In *Techniques in Protein Chemistry V*; Academic Press: San Diego, 1994; p 19.
(151) Bartlet-Jones, M.; Jeffery, W. A.; Hansen, H. F.; Pappin, D. J. C. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 737.
(152) Yates, J. R., III; McCormack, A. L.; Hayden, J. B.; Davey, M. P. *Cell Biology: A Laboratory Handbook*; Academic Press: New York, 1994; p 380.
(153) Takao, T.; Hori, H.; Okamoto, K.; Harada, A.; Kamachi, M.; Shimonishi, Y. *Rapid Commun. Mass Spectrom.* **1991**, *5*, 312.
(154) Schnolzer, M.; Jedrzejewski, P.; Lehmann, W. D. *Electrophoresis* **1996**, *17*, 945.
(155) Kosaka, T.; Takazawa, T.; Nakamura, T. *Anal. Chem.* **2000**, *72*, 1179.
(156) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015.
(157) Qin, J.; Herring, C. J.; Zhang, X. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 209.
(158) Huang, Z.-H.; Shen, T.; Wu, J.; Gage, D. A.; Watson, J. T. *Anal. Biochem.* **1999**, *268*, 305.
(159) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 7131.
(160) Yates, J. R., III. *Trends Genet.* **2000**, *16*, 5.
(161) Mortz, E.; O'Connor, P. B.; Roepstorff, P.; Kelleher, N. L.; Wood, T. D.; McLaffertey, F. W.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8264.
(162) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390.
(163) Hoving, S.; Munchbach, M.; Schmid, H.; Signor, L.; Lehmann, A.; Staudenmann, W.; Quadroni, M.; James, P. *Anal. Chem* **2000**, *72*, 1006.
(164) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327.
(165) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.
(166) Cox, A. L.; Skipper, J.; Chen, Y.; Henderson, R. A.; Darrow, T. L.; Shabanowitz, J.; Engelhard, V. H.; Hunt, D. F.; Slingluff, C. L. *Science* **1994**, *264*, 716.
(167) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426.
(168) Yates, J. R., III; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67*, 3202.
(169) Bruce, J. E.; Anderson, G. A.; Brands, M. D.; Pasa-Tolic, L.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 416.
(170) Solouki, T.; Marto, J. A.; White, F. M.; Guan, S.; Marshall, A. G. *Anal. Chem.* **1995**, *67*, 4139.
(171) Goodlett, D. R.; Bruce, J. E.; Smith, R. D.; Aebersold, R. Unpublished data from collaboration at B. M. I. Environmental and Molecular Sciences Laboratory in Richland, WA. 1998.

(172) O'Farrell, P. H. *J. Biol. Chem.* **1975**, *250*, 4007.
(173) Klose, J. *Humangentik* **1975**, *26*, 231.
(174) Gygi, S. P.; Rochon, Y.; Franza, B.; Aebersold, R. *Mol. Cell. Biol.* **1999**, *19*, 1729.
(175) Kurland, C. G. *FEBS Lett* **1991**, *285*, 165.
(176) Futcher, B. *Methods Cell Sci.* **1999**, *21*, 79.
(177) Perrot, M.; Sagliocco, F.; Mini, T.; Monribot, C.; Schneider, U.; Shevchenko, A.; Mann, M.; Jeno, P.; Boucherie, H. *Electrophoresis* **1999**, *20*, 2280.
(178) Corthals, G. L.; Wasinger, V. C.; Hochstrasser, D. F.; Sanchez, J. C. *Electrophoresis* **2000**, *21*, 1104.
(179) Washburn, M. P.; Yates, J. R. *Curr. Opin. Microbiol.* **2000**, *3*, 292.
(180) Tong, W.; Link, A.; Eng, J. K.; Yates, J. R., III. *Anal. Chem.* **1999**, *71*, 2270.
(181) Jensen, P. K.; Pasa-Tolic, L.; Peden, K. K.; Martinovic, S.; Lipton, M. S.; Anderson, G. A.; Tolic, N.; Wong, K. K.; Smith, R. D. *Electrophoresis* **2000**, *21*, 1372.
(182) Patterson, S. D. *Anal. Biochem.* **1994**, *221*, 1.
(183) Opiteck, G. J.; Jorgenson, J. W.; Anderegg, R. J. *Anal. Chem.* **1997**, *69*, 2283.
(184) Opiteck, G. J.; Lewis, K. C.; Jorgenson, J. W.; Anderegg, R. J. *Anal. Chem.* **1997**, *69*, 1518.
(185) Hsieh, Y. L.; Wang, H.; Elicone, C.; Mark, J.; Martin, S. A.; Regnier, F. *Anal. Chem.* **1996**, *68*, 455.
(186) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994.
(187) Kaufmann, R.; Kirsch, D.; Spengler, B. *Int. J. Mass Spectrom. Ion Processes* **1994**, *131*, 355.
(188) Masselon, C.; Anderson, G. A.; Harkewicz, R.; Bruce, J. E.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000**, *72*, 1918.
(189) DeRisis, L.; Iyer, V. R.; Brown, P. O. *Science* **1997**, *278*, 680.
(190) De Leenheer, A. P.; Thienpont, L. M. *Mass Spectrom. Rev.* **1992**, *11*, 249.
(191) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6591.
(192) Pasa-Tolic, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinovic, S.; Tolic, N.; Bruce, J. E.; Smith, R. D. *J. Am. Chem. Soc.* **1999**, *121*, 7949.
(193) Selle, H.; Schrader, M.; Schoeffski, P.; Hess, R.; Zucht, H.-D.; Heine, G.; Juergens, M.; Schulz-Knappe, P. Presented at the European Meeting on Biomarkers of Organ Damage and Dysfunction, Cambridge, U.K., April 3−7, 2000.
(194) Shrader, M. BioVisioN GmbH, Hannover, Germany, personal communication, 2000.
(195) Liu, Y.; Patricelli, M. P.; Cravatt, B. F. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14694.
(196) Simone, N. L.; Bonner, R. F.; Gillespie, J. W.; Emmert-Buck, M. R.; Liotta, L. A. *Trends Genet.* **1998**, *14*, 272.
(197) Simone, N. L.; Remaley, A. T.; Charboneau, L.; Petricoin, E. F., III; Glickman, J. W.; Emmert-Buck, M. R.; Fleisher, T. A.; Liotta, L. A. *Am. J. Pathol.* **2000**, *156*, 445.
(198) Lee, H.; Williams, S. K.; Giddings, J. C. *Anal. Chem.* **1998**, *70*, 2495.
(199) Rout, M. P.; Aitchison, J. D.; Suprapto, A.; Hjertaas, K.; Zhao, Y.; Chait, B. T. *J. Cell Biol.* **2000**, *148*, 635.
(200) Heller, M.; Goodlett, D. R.; Watts, J. D.; Aebersold, R. *Electrophoresis* **2000**, *21*, 2180.
(201) Bouveret, E.; Rigaut, G.; Shevchenko, A.; Wilm, M.; Seraphin B. *EMBO J.* **2000**, *19*, 1661.
(202) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Seraphin, B. *Nat. Biotechnol.* **1999**, *17*, 1030.
(203) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **2000**, *403*, 623.
(204) Krishna, R. G.; Wold, F. *PROTEINS: Analysis & Design*; Academic Press: San Diego, 1998; p 121.
(205) Charbonneau, H.; Tonks, N. K. *Annu. Rev. Cell Biol.* **1992**, *8*, 463.
(206) Fischer, E. H.; Krebs, E. G. *Biochim. Biophys. Acta* **1989**, *1000*, 297.
(207) Hunter, T. *Cell* **1987**, *50*, 823.
(208) Duclos, B.; Marcandier, S.; Cozzone, A. J. *Methods Enzymol.* **1991**, *201*, 10.
(209) Karin, M.; Hunter, T. *Curr. Biol.* **1995**, *5*, 747.
(210) Johnson, L. N.; Barford, D. *J. Biol. Chem.* **1996**, *265*, 2409.
(211) Joung, I.; Kim, T.; Stolz, L. A.; Payne, G.; Winkler, D. G.; Walsh, C. T.; Strominger, J. L.; Shin, J. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 5778.
(212) Watts, J. D.; Welham, M. J.; Kalt, L.; Schrader, J. W.; Aebersold, R. *J. Immunol.* **1993**, *151*, 6862.
(213) Pawson, T.; Nash, P. *Genes Dev.* **2000**, *14*, 1027.
(214) Wange, R. L.; Isakov, N.; Burke, T. R.; Otaka, A.; Roller, P. P.; Watts, J. D.; Aebersold, R.; Samelson, L. E. *J. Biol. Chem.* **1995**, *270*, 944.
(215) Watts, J. D.; Affolter, M.; Krebs, D. L.; Wange, R. L.; Samelson, L. E.; Aebersold, R. *J. Biol. Chem.* **1994**, *269*, 29520.
(216) Watts, J. D.; Brabb, T.; Bures, E. J.; Wange, R. L.; Samelson, L. E.; Aebersold, R. *FEBS Lett.* **1996**, *398*, 217.
(217) Haspel, R. L.; Darnell, J. E., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10188.
(218) Abu-Amer, Y.; Ross, F. P.; McHugh, K. P.; Livolsi, A.; Peyron, J. F.; Teitelbaum, S. L. *J. Biol. Chem.* **1998**, *273*, 29417.
(219) Garrison, T. R.; Zhang, Y.; Pausch, M.; Apanovitch, D.; Aebersold, R.; Dohlman, H. G. *J. Biol. Chem.* **1999**, *274*, 36387.
(220) Boyle, W. J.; van der Geer, P.; Hunter, T. *Methods Enzymol.* **1991**, *201*, 110.
(221) Martzen, M. R.; McCraith, S. M.; Spinelli, S. L.; Torres, F. M.; Fields, S.; Grayhack, E. J.; Phizicky, E. M. *Science* **1999**, *286*, 1153.
(222) Zhou, H.; Watts, J. D.; Aebersold, R. Presented at the 48th American Society for Mass Spectrometry Meeting, Long Beach, CA, June 2000.
(223) Weckwerth, W.; Willmitzer, L.; Fiehn, O. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1677.
(224) Gallis, B.; Corthals, G. L.; Goodlett, D. R.; Ueba, H.; Kim, F.; Presnell, S. R.; Figeys, D.; Harrison, D. G.; Berk, B. C.; Aebersold, R.; Corson, M. *J. Biol. Chem.* **1999**, *274*, 30101.
(225) Watts, J. D.; Krebs, D. L.; Wange, R. L.; Samelson, L. E.; Aebersold, R. *Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry*; American Chemical Society: Washington, DC, 1996; p 381.
(226) Yanagida, M.; Miura, Y.; Yagasaki, K.; Taoka, M.; Isobe, T.; Takahashi, N. *Electrophoresis* **2000**, *21*, 1890.
(227) Heffetz, D.; Fridkin, M.; Zick, Y. *Eur. J. Biochem.* **1989**, *182*, 343.
(228) Wu, J. J.; Yarwood, D. R.; Pham, Q.; Sills, M. A. *J. Biomol. Screening* **2000**, *5*, 23.
(229) Soskic, V.; Gorlach, M.; Poznanovic, S.; Boehmer, F. D.; Godovac-Zimmerman, J. *Biochemistry* **1999**, *38*, 1757.
(230) Goodlett, D. R.; Aebersold, R.; Watts, J. D *Rapid Commun. Mass Spectrom.* **2000**, *14*, 344.
(231) Verma, R.; Annan, R. S.; Huddleston, M. J.; Carr, S. A.; Reynard, G.; Deshaies, R. J. *Science* **1997**, *278*, 455.
(232) Gatti, A.; Traugh, J. A. *Anal. Biochem.* **1999**, *266*, 198.
(233) Porath, J. *Protein Expression Purif.* **1992**, *3*, 263.
(234) Boyle, W. J.; Smeal, T.; Defize, L. H.; Angel, P.; Woodgett, J. R.; Karin, M.; Hunter, T. *Cell* **1991**, *64*, 573.
(235) Affolter, M.; Watts, J. D.; Krebs, D. L.; Aebersold, R. *Anal. Biochem.* **1994**, *223*, 74.
(236) Neville, D. C.; Rozanas, C. R.; Price, E. M.; Gruis, D. B.; Verkman, A. S.; Townsend, R. R. *Protein Sci.* **1997**, *6*, 2436.
(237) Nuwaysir, L. M.; Stults, J. T. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 662.
(238) Tempst, P.; Link, A. J.; Riviere, L. R.; Fleming, M.; Elicone, C. *Electrophoresis* **1990**, *11*, 537.
(239) Katze, M. G.; Kwieciszewski, B.; Goodlett, D. R.; Blakely, C. M.; Nedderman, P.; Tan, S.-L.; Aebersold, R. *Virology* **2000**, *278*, 501.
(240) Annan R. S.; Carr S. A. *J. Protein Chem.* **1997**, *16*, 391.
(241) Covey, T. R.; Huang, E. C.; Henion, J. D. *Anal. Chem.* **1991**, *63*, 1193.
(242) Katta, V.; Chowdhury, S. K.; Chait, B. T. *Anal. Chem.* **1991**, *63*, 174.
(243) Cao, P.; Stults J. T. *J. Chromatogr. A* **1999**, *853*, 225.
(244) Meyer, H. E.; Eisermann, B.; Heber, M.; Hoffmann-Posorske, E.; Korte, H.; Weigt, C.; Wegner, A.; Hutton, T.; Donella-Deana, A.; Perich, J. W. *FASEB J.* **1993**, *7*, 776.
(245) Huddleston, M. J.; Annan, R. S.; Bean, M. F.; Carr, S. A. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 710.
(246) Hunter, A. P.; Games, D. E. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 559.
(247) Aebersold, R.; Figeys, D.; Gygi, S.; Corthals, G.; Haynes, P.; Rist, B.; Sherman, J.; Shang, Y.; Goodlett, D. *J. Protein Chem.* **1998**, *17*, 533.
(248) Gibson, B. W.; Cohen, P. *Methods Enzymol.* **1990**, *193*, 480.
(249) Carr S. A.; Huddleston M. J.; Annan, R. S. *Anal. Biochem.* **1996**, *239*, 180.
(250) Neubauer, G.; Mann, M. *Anal. Chem.* **1999**, *71*, 235.
(251) Annan, R. S.; Carr, S. A. *Anal. Chem.* **1996**, *68*, 3413.
(252) DeGnore, J. P.; Qin, J. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1175.
(253) Zhang, X.; Herring, C. J.; Romano, P. R.; Szczepanowska, J.; Brzeska, H.; Hinnebusch, A. G.; Qin, J. *Anal. Chem.* **1998**, *70*, 2050.
(254) Amankwa, L. N.; Harder, K.; Jirik, F.; Aebersold, R. *Protein Sci.* **1995**, *4*, 113.

CR990076H

articles

A proteomic view of the *Plasmodium falciparum* life cycle

Laurence Florens*, Michael P. Washburn†, J. Dale Raine‡, Robert M. Anthony§, Munira Grainger||, J. David Haynes§¶, J. Kathleen Moch§, Nemone Muster*, John B. Sacci§#, David L. Tabb*☆, Adam A. Witney§#, Dirk Wolters†#, Yimin Wu, Malcolm J. Gardner††, Anthony A. Holder||, Robert E. Sinden‡, John R. Yates*† & Daniel J. Carucci§**

*Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, California 92037, USA
†Department of Proteomics and Metabolomics, Torrey Mesa Research Institute, Syngenta Research & Technology, 3115 Merryfield Row, San Diego, California 92121-1125, USA
‡Infection and Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK
§Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40; and ¶ Department of Immunology, Walter Reed Army Institute of Research, Silver Spring, Maryland 20910-7500, USA
||The Division of Parasitology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
☆Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA
**Malaria Research and Reference Reagent Resource Center, American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209, USA
††The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

The completion of the *Plasmodium falciparum* clone 3D7 genome provides a basis on which to conduct comparative proteomics studies of this human pathogen. Here, we applied a high-throughput proteomics approach to identify new potential drug and vaccine targets and to better understand the biology of this complex protozoan parasite. We characterized four stages of the parasite life cycle (sporozoites, merozoites, trophozoites and gametocytes) by multidimensional protein identification technology. Functional profiling of over 2,400 proteins agreed with the physiology of each stage. Unexpectedly, the antigenically variant proteins of *var* and *rif* genes, defined as molecules on the surface of infected erythrocytes, were also largely expressed in sporozoites. The detection of chromosomal clusters encoding co-expressed proteins suggested a potential mechanism for controlling gene expression.

The life cycle of *Plasmodium* is extraordinarily complex, requiring specialized protein expression for life in both invertebrate and vertebrate host environments, for intracellular and extracellular survival, for invasion of multiple cell types, and for evasion of host immune responses. Interventional strategies including anti-malarial vaccines and drugs will be most effective if targeted at specific parasite life stages and/or specific proteins expressed at these stages. The genomes of *P. falciparum*[1] and *P. yoelii yoelii*[2] are now completed and offer the promise of identifying new and effective drug and vaccine targets.

Functional genomics has fundamentally changed the traditional gene-by-gene approach of the pre-genomic era by capitalizing on the success of genome sequencing efforts. DNA microarrays have been successfully used to study differential gene expression in the abundant blood stages of the *Plasmodium* parasite[3,4]. However, transcriptional analysis by DNA microarrays generally requires microgram quantities of RNA and has been restricted to stages that can be cultivated *in vitro*, limiting current large-scale gene expression analyses to the blood stages of *P. falciparum*. As several key stages of the parasite life cycle, in particular the pre-erythrocytic stages, are not readily accessible to study, and as differential gene expression is in fact a surrogate for protein expression, global proteomic analyses offer a unique means of determining not only protein expression, but also subcellular localization and post-translational modifications.

We report here a comprehensive view of the protein complements isolated from sporozoites (the infectious form injected by the mosquito), merozoites (the invasive stage of the erythrocytes), trophozoites (the form multiplying in erythrocytes), and gametocytes (sexual stages) of the human malaria parasite *P. falciparum*. These proteomes were analysed by multidimensional protein identification technology (MudPIT), which combines in-line, high-resolution liquid chromatography and tandem mass spectrometry[5]. Two levels of control were implemented to differentiate parasite from host proteins. By using combined host–parasite sequence databases and noninfected controls, 2,415 parasite proteins were confidently identified out of thousands of host proteins; that is, 46% of all gene products were detected in four stages of the *Plasmodium* life cycle (Supplementary Table 1).

Comparative proteomics throughout the life cycle

The sporozoite proteome appeared markedly different from the other stages (Table 1). Almost half (49%) of the sporozoite proteins

Table 1 **Comparative summary of the protein lists for each stage**

Protein count	Sporozoites	Merozoites	Trophozoites	Gametocytes
152	X	X	X	X
197	–	X	X	X
53	X	–	X	X
28	X	X	–	X
36	X	X	X	–
148	–	–	X	X
73	–	X	–	X
120	X	–	–	X
84	–	X	X	–
80	X	–	X	–
65	X	X	–	–
376	–	–	–	X
286	–	–	X	–
204	–	X	–	–
513	X	–	–	–
2,415	1,049	839	1,036	1,147

Whole-cell protein lysates were obtained from, on average, 17×10^6 sporozoites, 4.5×10^9 trophozoites, 2.75×10^9 merozoites, and 6.5×10^9 gametocytes.

¤ Present addresses: BRB 13-009, Department of Microbiology and Immunology, University of Maryland School of Medicine, 655 W. Baltimore St., Baltimore, Maryland 21201, USA (J.B.S.); Department of Medical Microbiology, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK (A.A.W.); and Ruhr-University Bochum, Institute of Analytical Chemistry, 44780 Bochum, Germany (D.W.).

were unique to this stage, which shared an average of 25% of its proteins with any other stage. On the other hand, trophozoites, merozoites and gametocytes had between 20% and 33% unique proteins, and they shared between 39% and 56% of their proteins. Consequently, only 152 proteins (6%) were common to all four stages. Those common proteins were mostly housekeeping proteins such as ribosomal proteins, transcription factors, histones and cytoskeletal proteins (Supplementary Table 1). Proteins were sorted into main functional classes based on the Munich Information Centre for Protein Sequences (MIPS) catalogue[6], with some adaptations for classes specific to the parasite, such as cell surface and apical organelle proteins (Fig. 1). When considering the annotated proteins in the database, some marked differences appeared between sporozoites and blood stages (Fig. 1). Although great care was taken to ensure that the results reflect the state of the parasite in the host, a portion of the data set may reflect the parasite's response to different purification treatments. However, the stage-specific detection of known protein markers at each stage established the relevance of our data set.

The merozoite proteome

Merozoites are released from an infected erythrocyte, and after a short period in the plasma, bind to and invade new erythrocytes. Proteins on the surface and in the apical organelles of the merozoite mediate cell recognition and invasion in an active process involving an actin-myosin motor. Four putative components of the invasion motor[7], merozoite cap protein-1 (MCP1), actin, myosin A, and myosin A tail domain interacting protein (MTIP), were abundant merozoite proteins (Supplementary Table 2). Abundant merozoite surface proteins (MSPs) such as MSP1 and MSP2 are linked by a glycosylphosphatidyl (GPI) anchor to the membrane, and both have been implicated in immune evasion (reviewed in ref. 8). A second family of peripheral membrane proteins, represented by MSP3 and MSP6, was also detected (Fig. 2a), although these proteins are largely soluble proteins of the parasitophorous vacuole, which are released on schizont rupture. Other vacuolar proteins, such as the acidic basic repeat antigen (ABRA) and serine repeat antigen (SERA), were detected in the merozoite fraction, but some such as S-antigen[9] were not (Supplementary Table 2). Notably, MSP8 and a related MSP8-like protein were only identified in sporozoites (Fig. 2a). Some MSPs are diverse in sequence and may be extensively modified by proteolysis; these features, together with the association of a variety of peripheral and soluble proteins, provide for a complex surface architecture.

Many apical organellar proteins, in the micronemes and rhoptries, have a single transmembrane domain. Among these proteins, apical membrane antigen 1 (AMA1) and MAEBL were found in both sporozoite and merozoite preparations (Fig. 2a). Erythrocyte-binding antigens (EBA), such as EBA 175 and EBA 140/BAEBL, were found only in the merozoite and trophozoite fractions. Of note, the reticulocyte-binding protein (PfRH) family (PFD0110w, MAL13P1.176, PF13_01998, PFL2520w and PFD1150c), which has similarity with the Py235 family of *P. y. yoelii* rhoptry proteins and the *Plasmodium vivax* reticulocyte-binding proteins, was not detected in the merozoite fraction. Some PfRH proteins were, however, detected in sporozoites (Fig. 2a), including RH3, which is a transcribed pseudogene in blood stages[10]. Components of the low molecular mass rhoptry complex, the rhoptry-associated proteins (RAP) 1, 2 and 3, were all found in merozoites. RAP1 was also detected in sporozoites. The high molecular mass rhoptry protein complex (RhopH), together with ring-infected erythrocyte surface antigen (RESA), which is a component of dense granules, is transferred intact to new erythrocytes at or after invasion and may contribute to the host cell remodelling process. RhopH1, RhopH2 (PFI1445w; Ling, I. T., *et al.*, unpublished data) and RhopH3 were found in the merozoite proteome. RhopH1 (PFC0120w/PFC0110w) has been shown to be a member of the cyto-adherence linked asexual gene family (CLAG)[11]; however, the presence of CLAG9 in the merozoite fraction (Fig. 2a) suggests that CLAG9 may also be a RhopH protein, casting some doubt on the proposed role for this protein in cyto-adherence[12].

The trophozoite proteome

After erythrocyte invasion the parasite modifies the host cell. The principal modifications during the initial trophozoite phase (lasting about 30 h) allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cyto-adherence, and to digest the cytoplasmic contents, particularly haemoglobin, in its food vacuole. In the next phase of schizogony (the final ~18 h of the asexual development in the blood cell), nuclear division is followed by merozoite formation and release.

Knob-associated histidine-rich protein (KAHRP) and erythrocyte membrane proteins 2 and 3 (EMP2 and -3) bind to the erythrocyte cytoskeleton (Fig. 2a). Of the proteins of the parasitophorous vacuole and the tubovesicular membrane structure extending into the cytoplasm of the red blood cell, three (the skeleton-binding protein 1, and exported proteins EXP1 and EXP2) were represented by peptides (Fig. 2a); although a fourth (Sar1 homologue, small GTP-binding protein; PFD0810w) was not. It is likely that one or more of the hypothetical proteins detected only in the trophozoite sample are involved in these unusual structures.

Digestion of haemoglobin is a major parasite catabolic process[13]. Members of the plasmepsin family (aspartic proteinases; PF14_0075 to PF14_0078)[14], falcipain family (cysteine proteinases; PF11_0161,

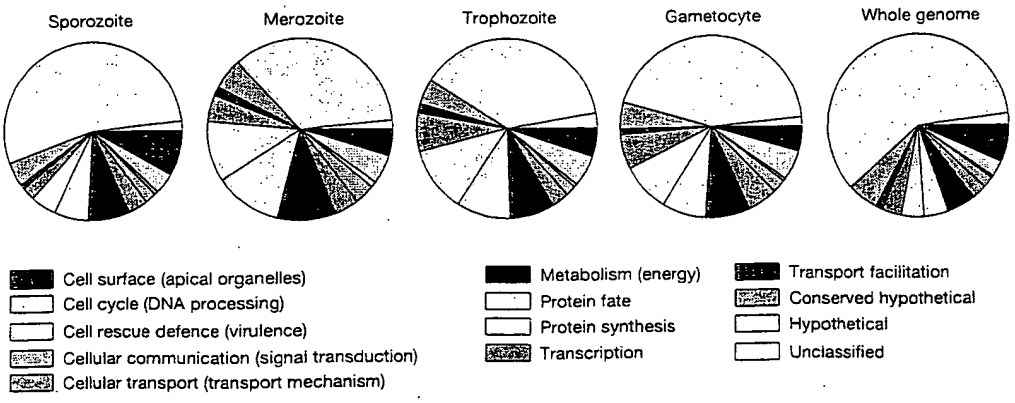

Figure 1 Functional profiles of expressed proteins. Proteins identified in each stage are plotted as a function of their broad functional classification as defined by the MIPS catalogue[6]. To avoid redundancy, only one class was assigned per protein. The complete protein list is given in Supplementary Table 1.

PF11_0162 and PF11_0165)[15], and falcilysin (a metallopeptidase; PF13_0322)[16] implicated in this process were all clearly identified (Supplementary Table 1). Several proteases expressed in the merozoite and trophozoite fractions, and not involved in haemoglobin digestion, may be important in parasite release at the end of schizogony, invasion of the new cell, or merozoite protein processing. Possible candidates for this mechanism include cysteine proteinases of the falcipain and SERA families, or subtilisins such as SUB1 and SUB2, both located in apical organelles (Fig. 2a).

The gametocyte proteome

Stage V gametocytes are dimorphic, with a male:female ratio of 1:4. They are arrested in the cell cycle until they enter the mosquito where development is induced within minutes to form the male and female gametes. Gametocyte structure reflects these ensuing fates; that is, the female has abundant ribosomes and endoplasmic reticulum/vesicular network to re-initiate translation, whereas the male is largely devoid of ribosomes and is terminally differentiated[17].

Gametocyte-specific transcription factors, RNA-binding proteins, and gametocyte-specific proteins involved in the regulation of messenger RNA processing (particularly splicing factors, RNA helicases, RNA-binding proteins, ribonucleoproteins (RNPs) and small nuclear ribonucleoprotein particles (snRNPS)) were highly represented in the gametocyte proteome (Supplementary Table 1). Transcription in the terminally differentiated gametocytes is 'suppressed', but the female gametocytes contain mRNAs encoding gamete/zygote/ookinete surface antigens (for example, P25/28)

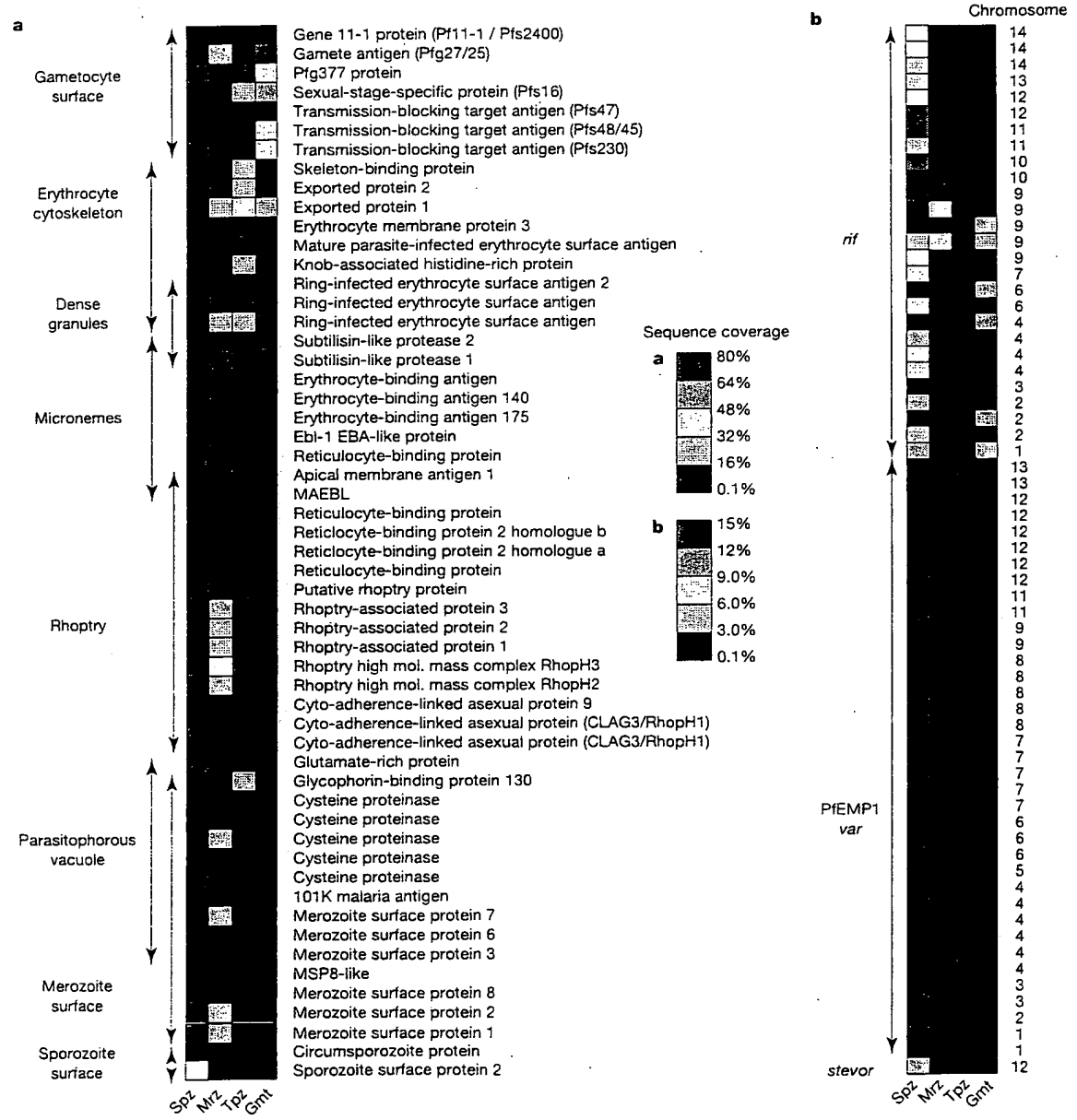

Figure 2 Expression patterns of known stage-specific proteins. **a**, Cell surface, organelle, and secreted proteins are plotted as a function of their known subcellular localization. **b**, *stevor*, *var* and *rif* polymorphic surface variants are plotted as a function of the chromosome encoding their genes. The matrices are colour-coded by sequence coverage measured in each stage (proteins not detected in a stage are represented by black squares). Locus names associated with these proteins are listed in Supplementary Table 2. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.

that are subject to post-transcriptional control; this control is released rapidly during gamete development[17]. Ribosomal proteins were largely represented: 82% of known small subunit (SSU) proteins and 69% of known large subunit (LSU) proteins were detected in gametocytes compared to 94% and 82%, respectively, from all stages examined (Supplementary Table 1). We suggest that this reflects the accumulation of ribosomes in the female gametocyte to accommodate for the sudden increase in protein synthesis required during gametogenesis and early zygote development.

Other protein groupings highly represented in the gametocyte were in the cell cycle/DNA processing and energy classes (Fig. 1). The former is consistent with the biological observation that the mature gametocyte is arrested in G0 of the cell cycle and will require a full complement of pre-existing cell cycle regulatory cascades to respond, within seconds, to the gametogenesis stimuli (that is, xanthurenic acid and a drop in temperature)[18]. Metabolic pathways of the malaria parasite may be stage-specific, with asexual blood stage parasites dependent on glycolysis and conversion of pyruvate to lactate (L-lactate dehydrogenase) for energy. In the gametocyte and sporozoite preparations, peptides from enzymes involved in the mitochondrial tricarboxylic acid (TCA) cycle and oxidative phosphorylation were identified (Table 2). This observation suggests that gametocytes have fully functional mitochondria as a pre-adaptation to life in the mosquito, as suggested by morphological and biochemical studies[19] and their sensitivity to anti-malarials attacking respiration (primaquine and artimesinin-based products)[17]. It will be interesting to observe whether other mosquito and liver stages, which show similar drug sensitivities, express the same metabolic proteome.

Cell surface proteins (Fig. 1) included most of the known surface antigens (Fig. 2a and Supplementary Table 2). However, Pfs35 and a sexual stage-specific kinase (PF13_0258) were not detected. Nevertheless the cultured gametocytes analysed in this study expressed a specific repertoire of rifin and PfEMP1 proteins (Fig. 2b and Supplementary Table 2). Together these observations suggest that the gametocyte, which is very long-lived in the red blood cell (that is, 9–12 days compared with 2 days for the pathogenic asexual parasites), expresses a limited repertoire of the highly polymorphic families of surface antigens so widely represented in the asexual parasites.

The sporozoite proteome

Sporozoites are injected by the mosquito during ingestion of a blood meal. Although, they are in the blood stream for only minutes, sporozoites probably require mechanisms to evade the host humoral immune system in order for at least a fraction of the thousands of sporozoites injected by the mosquito to survive the hostile environment in the blood and successfully invade hepatocytes.

The main class of annotated sporozoite proteins identified was cell surface and organelle proteins (Fig. 1). Sporozoites are an invasive stage and possess the apical complex machinery involved in host cell invasion. As observed in the analysis of the *P. y. yoelii* sporozoite transcriptome[20], actin and myosin were found in the motile sporozoites (Supplementary Table 2). Many proteins associated with rhoptry, micronemes and dense granules were detected (Fig. 2a). Among the proteins found were known markers of the sporozoite stage, such as the circumsporozoite protein (CSP) and sporozoite surface protein 2 (SSP2; also known as TRAP), both present in large quantities at the sporozoite surface (Fig. 2a). Peptides derived from CTRP (circumsporozoite protein and thrombospondin-related adhesive protein (TRAP)-related protein), an ookinete cell surface protein involved in recognition and/or motility[21], were detected in the sporozoite fractions (Supplementary Table 1).

Most surprisingly, peptides derived from multiple *var* (coding for PfEMP1) and *rif* genes were identified in the sporozoite samples. PfEMP1 and rifins are coded for by large multigene families (*var* and *rif*)[22,23] and are present on the surface of the infected red blood cell. No peptides derived from *rif* genes were identified in the trophozoite sample, whereas sporozoites expressed 21 different rifins and 25 PfEMP1 isoforms (Fig. 2b); that is, a total of 14% of the *rif* genes and 33% of the *var* genes encoded by the genome. Furthermore, very little overlap was observed between stages: only ten PfEMP1 and two rifin isoforms expressed in sporozoites were found in other stages. Whereas in the blood stream the asexual stage parasites undergo asexual multiplication and therefore have an opportunity to undergo antigenic 'switching' of the variant antigen genes, the non-replicative sporozoites may not have this opportunity. Expressing such a polymorphic array of *var* (PfEMP1) and *rif* genes could be part of a sporozoite survival mechanism.

Chromosomal clusters encoding co-expressed proteins

The distinct proteomes of each stage of the *Plasmodium* life cycle suggested that there is a highly coordinated expression of *Plasmodium* genes involved in common processes. Co-expression groups are a widespread phenomenon in eukaryotes, where mRNA array analyses have been used to establish gene expression profiles. Analysis of co-regulated gene groups facilitates both searching for regulatory motifs common to co-regulated genes, and predicting protein function on the basis of the 'guilt by association' model. Furthermore, mRNA analyses in *Saccharomyces cerevisiae*[24] and *Homo sapiens*[25,26] have demonstrated that co-regulated genes do not map to random locations in the genome but are in fact

Table 2 **Examples on enzymes in stage-specific metabolic pathways**

Locus	Stage				Enzyme	EC number†	Reaction catalysed
	Spz*	Mrz*	Tpz*	Gmt*			
End of glycolysis							
PF10_0363	1.2	–	2.4	–	Pyruvate kinase	2.7.1.40	*P*-enolpyruvate to pyruvate
MAL6P1.160	8.6	66.9	18.8	14.7	Pyruvate kinase		
PF13_0141	46.2	83.9	70.9	78.8	L-lactate dehydrogenase	1.1.1.27	Pyruvate to lactate
TCA cycle and oxidative phosphorylation							
PF10_0218	12.3	–	–	–	Citrate synthase	4.1.3.7	Acetyl coA + oxaloacetate to citrate
PF13_0242	3.2	–	16.9	8.8	Isocitrate dehydrogenase (NADP)	1.1.1.41	Isocitrate to 2-oxoglutarate + CO_2
PF08_0045	2.9	–	2.2	23.1	2-Oxoglutarate dehydrogenase e1 component	1.2.4.2	2-Oxoglutarate to succinyl CoA
PF10_0334	–	–	3.5	27.7	Flavoprotein subunit of succinate dehydrogenase	1.3.5.1	Succinate to fumarate
PFL0630w	3.7	–	–	12.1	Iron-sulphur subunit of succinate dehydrogenase		
PF14_0373	–	–	–	12.7	Ubiquinol cytochrome oxidoreductase	1.10.2.2.	Ubiquinol to cytochrome *c* reductase in electron transport
PFB0795w	–	–	–	14.2	ATP synthase F1, α-subunit		
PFI1365w	–	–	–	8.8	Cytochrome *c* oxidase subunit	1.9.3.1	
PFI1340w	–	–	–	8.8	Fumarate hydratase	4.2.1.2	Fumarate to malate
MAL6P1.242	30.4	–	–	40.9	Malate dehydrogenase	1.1.1.37	Malate to oxaloacetate

Plasmodium metabolic pathways can be found at http://www.sites.huji.ac.il/malaria/. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.
*The sequence coverage (that is, the percentage of the protein sequence covered by identified peptides) measured in each stage is reported.
†Enzyme Commission (EC) numbers are reported for each protein.

frequently organized into gene clusters on a chromosome. Gene clustering in *Plasmodium* species has been demonstrated. Ordered arrays of genes involved in virulence and antigenic variation (for example, *var*, *vir* and *rif* genes) are located in the subtelomeric regions of the chromosomes[27,28].

To determine whether gene clustering exists along the entire *P. falciparum* genome, genes whose protein products were detected in our analysis were mapped onto all 14 chromosomes in a stage-dependent manner (Fig. 3a). The 2,415 proteins identified represented an average of 45% of the open reading frames (ORFs) predicted per chromosome. The number of protein hits by chromosome was similar for all stages: sporozoite, merozoite, trophozoite and gametocyte protein lists constituting 19.7%, 15.8%, 19.5% and 21.6% of the predicted ORFs per chromosomes, respectively. Groups of three or more consecutive loci whose protein products were detected in a particular stage were defined as chromosomal clusters encoding co-expressed proteins (Fig. 3b). On the basis of this definition a total of 98 clusters containing 3 loci, 32 clusters containing 4 loci, 5 clusters containing 5 loci, and 3 clusters containing 6 loci were identified (Supplementary Table 3). For each chromosome, the frequency of finding clusters encoding co-expressed proteins containing 3–6 adjacent loci markedly exceeded the probability of finding such clusters by chance (see the footnote of Supplementary Table 3 for details on the probability calculation). Therefore, chromosomal clusters encoding co-expressed proteins were prevalent in the *P. falciparum* genome.

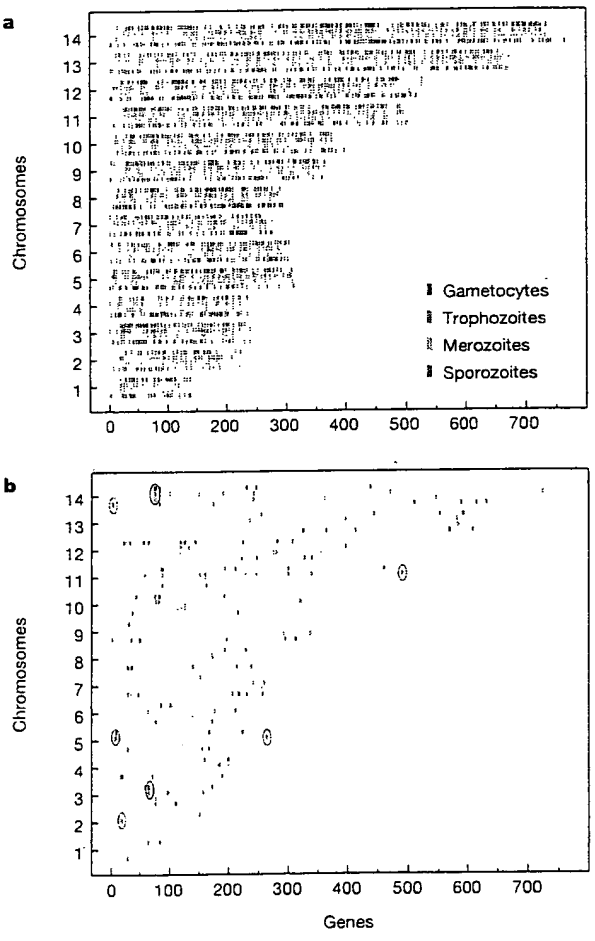

Figure 3 Distribution of expressed proteins by chromosome. **a**, For each stage, genes whose products were detected (coloured vertical bars) are plotted in the order they appear on their chromosome (grey boxes). **b**, Groups of at least three consecutive expressed genes are defined as chromosomal clusters of co-expressed proteins. Examples of such clusters, circled in **b**, are specified in Table 3 and the complete description of the 138 clusters can be found in Supplementary Table 3.

Functionally related genes have been shown to cluster in the *S. cerevisiae*[24] and human genomes[26]. This phenomenon also occurs in *P. falciparum*. A total of 138 clusters encoding co-expressed proteins were identified and 67 of them (49%) contained at least two loci that have been functionally annotated. Of these 67 clusters, 30 contained at least two loci whose annotation clearly indicates that the proteins are functionally related. For example, clusters on chromosomes 3, 5 and 10 contained ribosomal proteins, proteins involved in protein modification, and proteins involved in nucleotide metabolism, respectively (Table 3). Chromosome 14 contained a cluster of four aspartic proteases co-expressed in all of the blood stages (Table 3). This cluster was not detected in sporozoites, where no haemoglobin degradation is expected to occur. Interestingly, whereas the falcipain gene cluster on chromosome 11 appeared in our analysis as a cluster of co-expressed proteins (Supplementary Table 3), the SERA gene cluster on chromosome 2, coding for proteins that share a papain-like sequence motif[29], did not. Of the ten sporozoite-specific clusters, five involved *var* and *rif* genes, such as the *rif* cluster located in the subtelomeric domain of chromosome 14 (Table 3). On the basis of their presence in clusters encoding co-expressed proteins, we were able to suggest functional roles for 24 proteins annotated as hypothetical in the *P. falciparum* genome (Supplementary Table 3). For example, a gametocyte-specific cluster on chromosome 13 encoded two transmission-blocking antigens (Pfs48/45 and Pfs47) and a hypothetical protein, PF13_0246, which might be a gametocyte surface protein. Two clusters on chromosomes 2 and 11 were highly specific to the trophozoite stage (Table 3). Each of these clusters contained well-known secreted and surface proteins, namely KAHRP, PfEMP3, antigen 332, and RESA, all of which have been implicated in knob formation. The highly coordinated expression of these genes makes the three hypothetical proteins listed in these trophozoite-specific gene clusters possible candidates for involvement in cyto-adherence.

Discussion

Although sample handling is a principal consideration when studying pathogens, the expression of large numbers of previously identified proteins was consistent with their published expression profiles, validating our data set as a meaningful sampling of each stage's proteome. This is a particularly important aspect of our analysis as 65% of the 5,276 genes encoded by the *P. falciparum* genome are annotated as hypothetical[1], and of the 2,415 expressed proteins we identified, 51% are hypothetical proteins (Supplementary Table 1). Our results confirmed that these hypothetical ORFs predicted by gene modelling algorithms were indeed coding regions. Furthermore, from all four stages analysed, we identified 439 proteins predicted to have at least one transmembrane segment or a GPI addition signal (18% of the data set) and 304 soluble proteins with a signal sequence; that is, potentially secreted or located to organelles. Well over half of the secreted proteins and integral membrane proteins detected were annotated as hypothetical (Supplementary Table 4). The obvious interest in this class of proteins is that, with no homology to known proteins, they represent potential *Plasmodium*-specific proteins and may provide targets for new drug and vaccine development.

Our comprehensive large-scale analysis of protein expression showed that most surface proteins are more widely expressed than initially thought. In particular, the *var* and *rif* genes, which were thought to be involved in immune evasion only in the blood stage, have now been shown to be expressed in apparently large and varied numbers at the sporozoite stage. These surface proteins might be involved in general interaction processes with host cells and/or immune evasion. An alternative hypothesis is that stage-specific

Table 3 Examples of chromosomal gene clusters encoding co-expressed proteins

Chromosome	ID	Locus	Stage				Description	Class	SP	TM
			Spz	Mrz	Tpz	Gmt				
3	64	PFC0285c	2.1	12.7	33.2	18.7	T-complex protein β-subunit	Protein fate	0	0
3	65	PFC0290w	8.3	–	33.8	18.6	40S ribosomal protein S23	Protein synthesis	0	0
3	66	PFC0295c	–	14.9	52.5	21.3	40S ribosomal protein S12	Protein synthesis	0	0
3	67	PFC0300c	–	12.1	30.4	17.9	60S ribosomal protein L7	Protein synthesis	0	0
5	263	PFE1345c	–	–	1.9	1.6	Minichromosome maintenance protein 3	Cell transport	0	0
5	264	PFE1350c	–	–	22.4	–	Ubiquitin-conjugating enzyme	Protein fate	0	0
5	265	PFE1355	–	4.8	2.6	2.6	Ubiquitin carboxy-terminal hydrolase	Protein fate	0	0
5	266	PFE1360c	–	–	7.7	–	Methionine aminopeptidase	Protein fate	0	0
10	119	PF10_0121	10.8	74.5	29	–	Hypoxanthine phosphoribosyltransferase	Metabolism	0	0
10	120	PF10_0122	5.4	6.1	–	6.1	Phosphoglucomutase	Metabolism	0	0
10	121	PF10_0123	–	11.7	–	–	GMP synthetase	Metabolism	0	0
10	122	PF10_0124	0.9	1.8	–	–	Hypothetical protein		0	0
14	74	PF14_0074	26.6	–	–	4.9	Hypothetical protein		0	0
14	75	PF14_0075	–	26.5	43.2	47.4	Plasmepsin	Protein fate	1	0
14	76	PF14_0076	–	6.6	35.2	10	Plasmepsin 1	Protein fate	1	0
14	77	PF14_0077	–	21.2	43	11.5	Plasmepsin 2	Protein fate	1	0
14	78	PF14_0078	–	14.2	52.8	29.9	HAP protein	Protein fate	1	0
14	2	PF14_0002	3.5	–	–	–	Rifin	Surface or organelles	0	1
14	3	PF14_0003	7.9	–	–	–	Rifin	Surface or organelles	1	2
14	4	PF14_0004	6.5	–	–	–	Rifin	Surface or organelles	1	2
2	18	PFB0090c	–	–	3	–	Hypothetical protein, conserved		0	0
2	19	PFB0095c	–	–	3.4	–	Erythrocyte membrane protein 3	Surface or organelles	1	0
2	20	PFB0100c	–	1.5	24.8	–	Knob-associated histidine-rich protein	Surface or organelles	1	0
11	489	PF11_0506	–	–	6.3	4.4	Hypothetical protein		0	1
11	490	PF11_0507	–	–	0.8	–	Antigen 332	Surface or organelles	0	0
11	491	PF11_0508	–	–	3.3	–	Hypothetical protein		0	0
11	492	PF11_0509	–	6.4	3	–	RESA	Surface or organelles	0	0
13	443	PF13_0246	4.5	–	–	8.6	Hypothetical protein		0	0
13	444	PF13_0247	–	–	–	32.4	Transmission-blocking target antigen precursor (Pfs48/45)	Surface or organelles	1	1
13	445	PF13_0248	–	–	–	7.1	Transmission-blocking target antigen precursor (Pfs47)	Surface or organelles	1	1

Clusters of at least three consecutive genes encoding co-expressed proteins are reported with their position (ID) on the chromosome, the sequence coverage measured for these proteins in each stage (%), their current annotation and functional class, and the predicted presence of signal peptide (SP) or transmembrane domains (TM) (based on the TMHMM[43], a transmembrane (TM) helices prediction method based on a hidden Markov model (HMM), big-PI Predictor[44] and SignalP[45] algorithms).

regulation is not as exact as previously thought.

One mechanism of protein expression control that contributes to stage specificity in *P. falciparum* arises from the chromosomal clustering of genes encoding co-expressed proteins. The clusters described in this study demonstrate a widespread high order of chromosomal organization in *P. falciparum* and probably correspond to regions of open chromatin allowing for co-regulated gene expression. The high (A + T) content of the *P. falciparum* genome makes the identification of regulatory sequences such as promoters and enhancers challenging[31,32]. Focusing analyses on stage-specific and multi-stage clusters will facilitate finding stage-specific and general *cis*-acting sequences in the *Plasmodium* genome and will help decipher gene expression regulation during the parasite life cycle.

The malaria parasite is a complex multi-stage organism, which has co-evolved in mosquitoes and vertebrates for millions of years. Designing drugs or vaccines that substantially and persistently interrupt the life cycle of this complex parasite will require a comprehensive understanding of its biology. The *P. falciparum* genome sequence and comparative proteomics approaches may initiate new strategies for controlling the devastating disease caused by this parasite. □

Methods

Parasite material

Plasmodium falciparum clone 3D7 (Oxford) was used throughout. Sporozoites were initially isolated from the salivary glands of *Anopheles stephansi* mosquitoes, 14 days after infection, by centrifugation in a Renografin 60 gradient, as described[33]. Four sporozoite samples were used as is. A fifth sample underwent an additional purification step on Dynabeads M-450 Epoxy coupled to NFS1 (an anti-*P. falciparum* CS protein monoclonal antibody)[34] according to the manufacturer's instructions (Dynal). Trophozoite-infected erythrocytes from synchronized cultures were purified on 70% Percoll-alanine[30], and the trophozoites released from the erythrocytes[35]. Of the of 260 parasitized erythrocytes counted by Giemsa-stained thin-blood film, 100% were identified as trophozoites. Merozoites were prepared essentially as described in ref. 36, using highly synchronized schizonts and purifying the merozoites by passage through membrane filters. Starting with synchronized asexual parasites grown in suspension culture as described[37,38], gametocytes were prepared by daily media changes of static cultures at 37 °C. When there were very few mature asexual stages present, gametocyte-infected erythrocytes were collected from the 52.5%/45% and 45%/30% interfaces of a Percoll gradient[39]. The gametocytes consisted mostly of stage IV and V parasites with minor contamination (<3%) from mixed asexual stage parasites. Finally, cellular debris from the upper bodies of parasite-free *A. stephansi* and non-infected human erythrocytes were used as controls for sporozoites and blood-stage parasites, respectively. Every effort was made to minimize enzymatic activity and protein degradation during sampling, and the subsequent isolation of the parasites; however, we cannot exclude that some of the differences in protein profiles that we observe between the different life-cycle stages may be a consequence of the sample-handling procedures.

Cell lysis

Five sporozoite, four merozoite, four trophozoite and three gametocyte preparations were lysed, digested and analysed independently. Cell pellets were first diluted ten times in 100 mM Tris-HCl pH 8.5, and incubated in ice for 1 h. After centrifugation at 18,000 *g* for 30 min, supernatants were set aside and microsomal membrane pellets were washed in 0.1 M sodium carbonate, pH 11.6. Soluble and insoluble protein fractions were separated by centrifugation at 18,000 *g* for 30 min. Supernatants obtained from both centrifugation steps were either combined (sporozoites, trophozoites and merozoites) or digested and analysed independently (gametocytes).

Peptide generation and analysis

The method follows that of Washburn *et al.*[5], with the exception that Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl; Pierce) was used to reduce urea-denatured proteins. Peptide mixtures were analysed through MudPIT as described[5].

Protein sequence databases

The *P. falciparum* database contained 5,283 protein sequences. Spectra resulting from contaminant mosquito and erythrocyte peptides had to be taken into account in the sporozoite and blood-stage samples, respectively. Tandem mass spectrometry (MS/MS) data sets from blood stages were therefore searched against a database containing both *P. falciparum* protein sequences and 24,006 ORFs from the human, mouse and rat RefSeq NCBI databases. At the date of the searches, the *Anopheles gambiae* genome was not available. The NCBI database contained 922 *Anopheles* and 313 *Aedes* proteins, which were combined to the 14,335 ORFs of the NCBI *Drosophila melanogaster*[40] database to create a control diptera database. Finally, these databases were complemented with a set of 172 known protein contaminants, such as proteases, bovine serum albumin and human keratins.

MS/MS data set analysis

The SEQUEST algorithm was used to match MS/MS spectra to peptides in the sequence databases[41]. To account for carboxyamidomethylation, MS/MS data sets were searched with a relative molecular mass of 57,000 (M_r 57K) added to the average molecular mass of cysteines. Peptide hits were filtered and sorted with DTASelect[42]. Spectra/peptide matches were only retained if they were at least half-tryptic (Lys or Arg at either end of the identified peptide) and with minimum cross-correlation scores (XCorr) of 1.8 for +1, 2.5 for +2, and 3.5 for +3 spectra and DeltaCn (top match's XCorr minus the second-best match's XCorr divided by the top match's XCorr) of 0.08. Peptide hits were deemed unambiguous only if they were not found in non-infected controls and were uniquely assigned to parasite proteins by searching against combined parasite–host databases. Finally, for low coverage loci, peptide/spectrum matches were visually assessed on two main criteria: any given MS/MS spectrum had to be clearly above the baseline noise, and both *b* and *y* ion series had to show continuity. The Contrast tool[42] was used to compare and merge protein lists from replicate sample runs and to compare the proteomes established for the four stages.

Received 31 July; accepted 9 September 2002; doi:10.1038/nature01107.

1. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511 (2002).
2. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* 419, 512–519 (2002).
3. Ben Mamoun, C. *et al.* Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* 39, 26–36 (2001).
4. Hayward, R. E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* 35, 6–14 (2000).
5. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* 19, 242–247 (2001).
6. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34 (2002).
7. Pinder, J. C. *et al.* Actomyosin motor in the merozoite of the malaria parasite, *Plasmodium falciparum*: implications for red cell invasion. *J. Cell Sci.* 111, 1831–1839 (1998).
8. Holder, A. A. *Malaria Vaccine Development: a Multi-immune Response and Multi-stage Perspective* (ed. Hoffman, S. L.) 77–104 (ASM Press, Washington, 1996).
9. Coppel, R. L. *et al.* Isolate-specific S-antigen of *Plasmodium falciparum* contains a repeated sequence of eleven amino acids. *Nature* 306, 751–756 (1983).
10. Taylor, H. M. *et al. Plasmodium falciparum* homologue of the genes for *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins, which is transcribed but not translated. *Infect. Immun.* 69, 3635–3645 (2001).
11. Kaneko, O. *et al.* The high molecular mass rhoptry protein, RhopH1, is encoded by members of the clag multigene family in *Plasmodium falciparum* and *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* 118, 223–231 (2001).
12. Trenholme, K. R. *et al.* clag9: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc. Natl Acad. Sci. USA* 97, 4029–4033 (2000).
13. Klemba, M. & Goldberg, D. E. Biological roles of proteases in parasitic protozoa. *Annu. Rev. Biochem.* 71, 275–305 (2002).
14. Banerjee, R. *et al.* Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proc. Natl Acad. Sci. USA* 99, 990–995 (2002).
15. Rosenthal, P. J., Sijwali, P. S., Singh, A. & Shenai, B. R. Cysteine proteases of malaria parasites: targets for chemotherapy. *Curr. Pharm. Des.* 8, 1659–1672 (2002).
16. Eggleson, K. K., Duffin, K. L. & Goldberg, D. E. Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* 274, 32411–32417 (1999).
17. Sinden, R. E., Butcher, G. A., Billker, O. & Fleck, S. L. Regulation of infectivity of *Plasmodium* to the mosquito vector. *Adv. Parasitol.* 38, 53–117 (1996).
18. Billker, O., Shaw, M. K., Margo, G. & Sinden, R. E. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature* 392, 289–292 (1998).
19. Krungkrai, J., Prapunwattana, P. & Krungkrai, S. R. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. *Parasite* 7, 19–26 (2000).
20. Kappe, S. H. *et al.* Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl Acad. Sci. USA* 98, 9895–9900 (2001).
21. Dessens, J. T. *et al.* CTRP is essential for mosquito infection by malaria ookinetes. *EMBO J.* 18, 6221–6227 (1999).
22. Deitsch, K. W. & Wellems, T. E. Membrane modifications in erythrocytes parasitized by *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* 76, 1–10 (1996).
23. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* 96, 9333–9338 (1999).
24. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* 26, 183–186 (2000).
25. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292 (2001).
26. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* 31, 180–183 (2002).
27. Hernandez-Rivas, R. *et al.* Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol. Cell Biol.* 17, 604–611 (1997).
28. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* 410, 839–842 (2001).
29. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* 282, 1126–1132 (1998).
30. Kanaani, J. & Ginsburg, H. Metabolic interconnection between the human malarial parasite *Plasmodium falciparum* and its host erythrocyte. *J. Biol. Chem.* 264, 3194–3199 (1989).
31. Dechering, K. J. *et al.* Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell Biol.* 19, 967–978 (1999).
32. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* 405, 827–836 (2000).
33. Pacheco, N. D., Strome, C. P., Mitchell, F., Bawden, M. P. & Beaudoin, R. L. Rapid, large-scale isolation of *Plasmodium berghei* sporozoites from infected mosquitoes. *J. Parasitol.* 65, 414–417 (1979).
34. Mellouk, S. *et al.* Evaluation of an *in vitro* assay aimed at measuring protective antibodies against sporozoites. *Bull. World Health Organ.* 68 Suppl., 52–59 (1990).
35. Rabilloud, T. *et al.* Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis* 20, 3603–3610 (1999).
36. Blackman, M. J. Purification of *Plasmodium falciparum* merozoites for analysis of the processing of merozoite surface protein-1. *Methods Cell Biol.* 45, 213–220 (1994).
37. Haynes, J. D. & Moch, J. K. Automated synchronization of *Plasmodium falciparum* parasites by culture in a temperature-cycling incubator. *Methods Mol. Med.* 72, 489–497 (2002).
38. Haynes, J. D., Moch, J. K. & Smoot, D. S. Erythrocytic malaria growth or invasion inhibition assays with emphasis on suspension culture GIA. *Methods Mol. Med.* 72, 535–554 (2002).
39. Carter, R., Ranford-Cartwright, L. & Alano, P. The culture and preparation of gametocytes of *Plasmodium falciparum* for immunochemical, molecular, and mosquito infectivity studies. *Methods Mol. Biol.* 21, 67–88 (1993).
40. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195 (2000).
41. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989 (1994).
42. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 1, 21–26 (2002).
43. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580 (2001).
44. Eisenhaber, B., Bork, P. & Eisenhaber, F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* 11, 1155–1161 (1998).
45. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements

We are grateful to J. Graumann, R. Sadygov, G. Chukkapalli, A. Majumdar and R. Sinkovits for computer programming; C. Deciu for the probability calculations; and C. Delahunty and C. Vieille for critical reading of the manuscript. The authors acknowledge the support of the Office of Naval Research, the US Army Medical Research and Material Command, and the National Institutes of Health (to J.R.Y.). J.D.R. is funded by a Wellcome Trust Prize Studentship. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* clone 3D7 public before publication of the completed sequence. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.Y. (e-mail: jyates@scripps.edu).

Anal. Chem. 2004, *76*, 1134−1144

Ultra-High-Efficiency Strong Cation Exchange LC/RPLC/MS/MS for High Dynamic Range Characterization of the Human Plasma Proteome

Yufeng Shen, Jon M. Jacobs, David G. Camp, II, Ruihua Fang, Ronald J. Moore, and Richard D. Smith*

Biological Science Division, Pacific Northwest National Laboratory, Richland, Washington 99352

Wenzhong Xiao and Ronald W. Davis

Stanford Genome Technology Center, Stanford University School of Medicine, Palo Alto, California 94304

Ronald G. Tompkins

Department of Surgery, Shriners Burn Center and Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114

High-efficiency nanoscale reversed-phase liquid chromatography (chromatographic peak capacities of ~1000: Shen, Y.; Zhao, R.; Berger, S. J.; Anderson, G. A.; Rodriguez, N.; Smith, R. D. *Anal. Chem.* 2002, *74*, 4235. Shen, Y.; Moore, R. J.; Zhao, R.; Blonder, J.; Auberry, D. L.; Masselon, C.; Pasa-Tolic, L.; Hixson, K. K.; Auberry, K. J.; Smith, R. D. *Anal. Chem.* 2003, *75*, 3596.) and strong cation exchange LC was used to obtain ultra-high-efficiency separations (combined chromatographic peak capacities of $>10^4$) in conjunction with tandem mass spectrometry (MS/MS) for characterization of the human plasma proteome. Using conservative SEQUEST peptide identification criteria (i.e., without considering chymotryptic or elastic peptides) and peptide LC normalized elution time constraints, the separation quality enabled the identification of proteins over a dynamic range of greater than 8 orders of magnitude in relative abundance using ion trap MS/MS instrumentation. Between 800 and 1682 human proteins were identified, depending on the criteria used for identification, from a total of 365 μg of human plasma. The analyses identified relatively low-level (~pg/mL) proteins (e.g., cytokines) coexisting with high-abundance proteins (e.g., mg/mL-level serum albumin).

Tandem mass spectrometry (MS/MS) provides a basis for identifying proteins from protein mixtures.[1,2] Complex mixtures of proteins can be addressed if separations are also applied, with the addressable level of complexity increasing with separation quality. When conventional, moderately efficient (separation peak capacities of $\sim10^2$) reversed-phase liquid chromatography (RPLC) is used for proteomics separations, sample prefractionation (typically with strong cation exchange liquid chromatography, SCXLC) prior to RPLC/MS/MS has provided total (i.e., two-dimensional) separation peak capacities of up to $\sim10^3$. These separations have extended the achievable protein identification coverage[1−4] and the range of relative protein abundances detected (i.e., the dynamic range of the measurements) to ~4 orders of magnitude.[2] This protein identification capability has been sufficient for studying microbial proteomes and even some eukaryotic systems (e.g., a significant fraction of the yeast proteome is expressed over this dynamic range).[5] However, this dynamic range is generally insufficient for mammalian proteomics due to both an increase in proteomic complexity and an inherently broader dynamic range of interest.[6−9]

Broad characterization of the human plasma proteome may provide one of the greatest challenges, since it can contain low-level proteins from essentially any tissue type as well as other important protein classes (e.g., cytokines at pg/mL level) in the presence of a relatively few very abundant proteins (particularly serum albumin at 35−55 mg/mL).[9] Some techniques (e.g., immunoassays) can selectively detect specific proteins at a very low abundance but do not provide a basis for broad protein identification. Until 2002, the cumulative count of human plasma proteins was limited to ~300 proteins.[9] The recent application of an SCXLC/RPLC/MS/MS approach resulted in the identification of ~490 proteins.[10] However, this relative success was based upon

(1) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* 2001, *19*, 242−247.
(2) Wolters, D. A.; Washburn, M. P.; Yates, J. R. *Anal. Chem.* 2001, *73*, 5683−5690.
(3) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* 2003, *2*, 43−50.
(4) Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S. *Anal. Chem.* 2003, *75*, 3145−3152.
(5) Ghaemmaghami, S.; Huh, W.-K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* 2003, *425*, 737−741.
(6) O'Donovan, C.; Apweiler, R.; Bairoch, A. *Trends Biotechnol.* 2001, *19*, 178−181.
(7) Corthals, G. L.; Wasinger, V. C.; Hochstrasser, D. F.; Sanchez, J.-C. *Electrophoresis* 2000, *21*, 1104−1115.
(8) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Paša-Tolić, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. *Proteomics* 2002, *2*, 513−523.
(9) Anderson, N. L.; Anderson, N. G. *Mol. Cel. Proteomics* 2002, *1*, 845−867.

10.1021/ac034869m CCC: $27.50
Published on Web 01/13/2004

immunoglobulin depletion, an approach that is potentially problematic for quantitative measurements due to the variable and selective losses of other proteins along with the immunoglobulins.[10]

The dynamic range and coverage that results from MS/MS proteome analysis is a function of both the quality of the separation(s) applied and the MS platform. Our previous work has shown that RPLC on both capillary and nanoscale levels (150–15-μm column i.d.) can provide roughly equivalent separation powers (i.e., chromatographic peak capacities of $\sim 10^3$ when interfaced with MS)[11,12] as obtained using conventional 2-D LC/LC.[2] The experimental results have also demonstrated that high-efficiency RPLC separations can provide protein MS/MS identification coverage and measurement dynamic range similar to conventional 2-D SCXLC/RPLC separations ($\sim 10^4$).[13,14] Thus, implementation of such high-efficiency RPLC separations as part of a 2-D SCXLC/RPLC methodology should provide significantly greater overall efficiencies (i.e., chromatographic peak capacities of $> 10^4$) and should substantially extend the MS/MS protein measurement coverage and dynamic range.

In this study, we have evaluated ultra-high-efficiency SCXLC/RPLC/MS/MS separations for the extension of the dynamic range and coverage for the identification of human plasma proteins. We show that a protein identification dynamic range of $> 10^8$ can be achieved using conventional ion trap MS/MS instrumentation. This approach has resulted in the identification of >800 human plasma proteins from ~365 μg (or ~5 μL) of plasma without the need for depletion of high-abundant serum albumin or immunoglobulins.

EXPERIMENTAL SECTION

Nanoscale High-Efficiency RPLC and Capillary SCXLC Experiments. A previously described on-line micro-solid-phase extraction (microSPE)-nanoLC system,[13] with an 85 cm × 30 μm i.d. capillary (packed with 3-μm C18 particles, 300-Å surface pore size, Phenomenex, Torrance, CA) as the nanoLC column and a 4 cm × 75 μm i.d. capillary (containing 5-μm C18 particles, 300-Å surface pore size, Phenomenex) as the microSPE column, was used for the high-efficiency RPLC experiments. The sample was loaded onto the microSPE column in 10 μL of solution (~2 min for loading) and was switched on-line to the nanoLC column for separation with a mobile-phase gradient from 100% A (H_2O/acetic acid/trifluoroacetic acid, TFA, 100:0.2:0.1, v/v/v) to 70% B (acetonitrile/H_2O/TFA, 90:10:0.1, v/v/v) in 300 min. Both the microSPE and nanoLC operations were performed at 10 000 psi, and the mobile-phase components were HPLC-grade purchased from Aldrich (Milwaukee, WI). The separation efficiency and reproducibility of this nanoscale LC instrumentation has been previously evaluated and reported.[12,13]

For the capillary SCXLC separations, highly hydrophilic polysulfoethyl aspartamide-bonded silica particles (3-μm diameter, 300-Å pore size, PolyLC Inc.) were used for the stationary phase. These particles were packed into an 80 cm × 320 μm i.d. fused-silica capillary at 10 000 psi using 0.4 M phosphate buffer (pH 4) as the slurry packing solvent (the use of a high-concentration phosphate buffer was for conditioning the SCX column during packing and subsequent depressurization, i.e., ~15 h). The column was then washed for another 15 h with 4 mM phosphate buffer (pH 2.5). The SCXLC separations were completed by loading 10 μL of sample solution with a switching valve (Valco, Houston, TX) and gradient elution from mobile phase 100% A (4 mM phosphate buffer, pH 2.5) to 70% B (0.4 M phosphate buffer, pH 2.5) in ~200 min using two LC pumps (Isco, Lincoln, NE) operated at 10 000 psi. The separations were monitored using a UV detector at 215 nm (Spectra 100, Spectra-Physics, San Jose, CA) for sample fractionation. The column was passivated prior to sample fractionation, with a small amount of plasma sample (~1.5 μg) to shield possible stationary-phase active surface sites.

ESI MS/MS Experiments. A replaceable emitter (~5-μm-i.d. orifice tapered from a 15-μm-i.d. fused-silica capillary) was connected to the 30-μm-i.d RPLC column outlet using a home-manufactured union (20-μm internal pore) for ESI.[12] A Finnigan LCQ XP ion trap mass spectrometer (ThermoQuest Corp., San Jose, CA) was used for MS/MS experiments with a heated capillary temperature of 150 °C and an ESI voltage of 1.6 kV. An ion trap collision energy setting of 45% was applied for ion fragmentation, and data-dependent peak selection was used for analyses.

Data Analyses. The SEQUEST program (ThermoQuest Corp.) was used for peptide and protein identification by searching against the NCI Frederick ABCC nonredundant database containing 76 402 FASTA entries.[15] Peptide MS/MS assignments were filtered according to the criteria previously reported.[1] Peptide identifications required the following criteria: $X_{corr} > 1.9$ with charge state 1+ and full tryptic cleavage, $X_{corr} > 2.2$ with charge state 2+ and full or partial tryptic cleavage, $X_{corr} > 3$ with charge state 2+ for all peptides, or $X_{corr} > 3.75$ with charge state 3+ and full or partial tryptic cleavage. In addition, two different ΔCn cutoff values were used, >0.05 and >0.1, thus providing a set of identifications with different levels of confidence (Table 3). Also given are the proteins identified based upon other previously published criteria.[10,16] Additionally, these sets of identified proteins were further culled using a peptide normalized elution time (NET) criterion. This involved screening the peptides to eliminate those having significant differences in their observed versus calculated NET values obtained by the use of an artificial neural network that we have described previously.[17] Only peptides that had an observed NET that agreed within ±10% of the predicted NET were retained. Furthermore, a more stringent definition of "partial tryptic" peptides has been applied based upon previous reports[18–20] and our own experience to reduce false positives. This criterion

(10) Adkins, J. H.; Varnum, S. M.; Auberry, K. J.; Moore, R. J.; Angell, N. H.; Smith, R. D.; Springer, D. L.; Pounds, J. G. *Mol. Cell. Proteomics* **2002**, *1*, 947–955.

(11) Shen, Y.; Zhao, R.; Belov, M. E.; Conrads, T. P.; Anderson, G. A.; Tang, K.; Pasa-Tolic, L.; Veenstra, T. D.; Lipton, M. S.; Udseth, H.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 1766–1775.

(12) Shen, Y.; Zhao, R.; Berger, S. J.; Anderson, G. A.; Rodriguez, N.; Smith, R. D. *Anal. Chem.* **2002**, *74*, 4235–4249.

(13) Shen, Y.; Moore, R. J.; Zhao, R.; Blonder, J.; Auberry, D. L.; Masselon, C.; Pasa-Tolic, L.; Hixson, K. K.; Auberry, K. J.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 3596–3605.

(14) Shen, Y.; Tolic, N.; Masselon, C.; Pasa-Tolic, L.; Camp, D. G.; Hixson, K. K.; Zhao, R.; Anderson, G. A.; Smith, R. D. *Anal. Chem.*, in press.

(15) At: ftp://ftp.ncifcrf.gov.

(16) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, K. J.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* **2002**, *419*, 520–526.

(17) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pas-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 1039–1048.

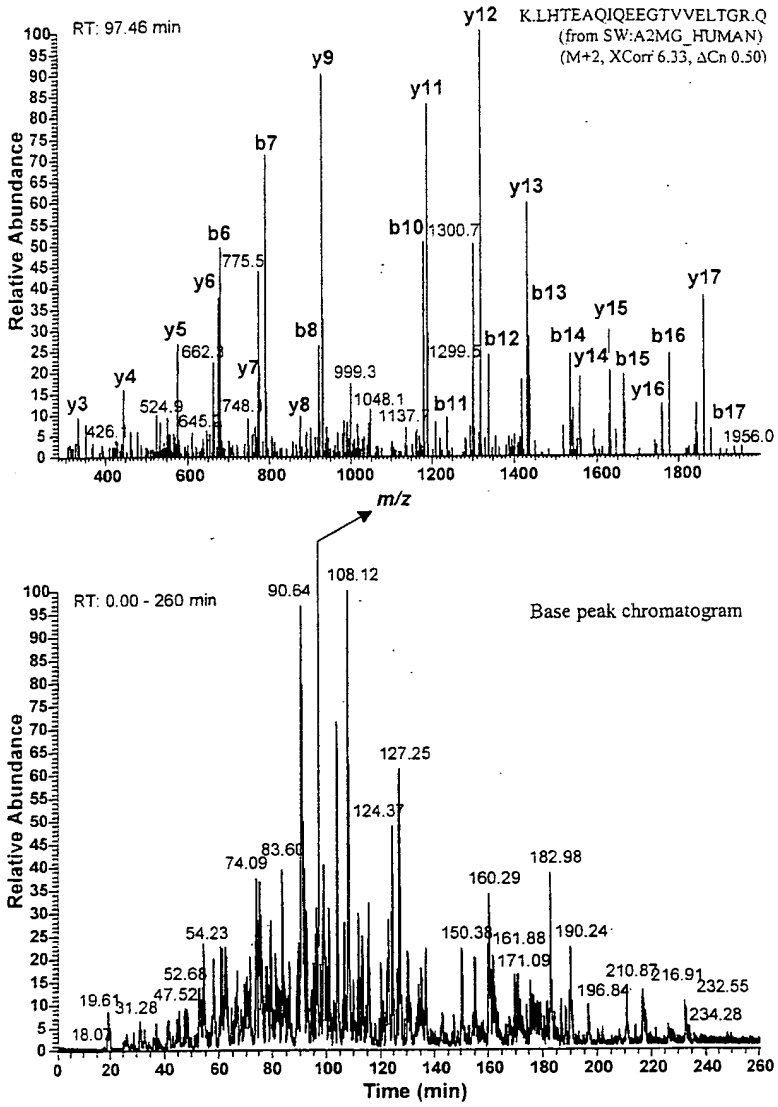

Figure 1. High-efficiency nanoscale RPLC/ion trap MS/MS of 0.5 μg of nondepleted human plasma tryptic digest. In the base peak chromatogram, highly intensive chromatographic peaks were observed at 90−130 min elution time for this nondepleted plasma sample; most low-to-moderate intensity peaks yielded high-quality MS/MS spectra, as shown, for peptide assignments. The RPLC and ion trap MS/MS experimental conditions are described in the Experimental Section.

requires that a partial tryptic peptide contain either K or R at one terminus of the peptide and V, E, F, L, or Y at the other terminus. The result of these steps, summarized in Table 3, is a set of identifications that are of high confidence. The complete set of identified peptides with SEQUEST X_{corr}, ΔCn scores is given in the Supporting Information.

Plasma Sample Preparation. The human blood plasma sample was obtained from Stanford University School of Medicine (Palo Alto, CA). An initial protein concentration of 65 mg/mL of plasma was determined using BCA Protein Assay (Pierce, Rockford, IL) after which the sample was diluted to 10 mg/mL for denaturation (9 M urea) and reduction (5 mM DTT). The sample was passed through a PD-10 desalting column (Amersham Pharmacia Biotech, Uppsala, Sweden), and the eluted protein was enzymatically digested using sequencing-grade modified porcine trypsin (Promega, Madison, WI) at a ratio of 1:50 (w/w, trypsin to protein) as instructed by the manufacturer. The digest was "cleaned" using a LC-18 SPE column (Supelco, Bellefonte, PA), after which the eluted peptides were lyophilized and reconstituted at a concentration of 15 μg/μL in 50 mM NH_4HCO_3 for further analysis.

RESULTS AND DISCUSSION

Human Plasma Proteomic Analysis Using Single-Dimension High-Efficiency RPLC/MS/MS. Figure 1 shows a reconstructed chromatogram and an MS/MS spectrum from a RPLC/MS/MS analysis of 0.5 μg of the plasma tryptic digest. Through database searching using SEQUEST, 110 proteins were identified (simply referred to as IDs in this study) from the assignment of

(18) Brandon, C.; Tooze, J. *Introduction to Protein Structure*; Garland Publishing: New York, 1991.
(19) Smith, R. L.; Shaw, E. *J. Biol. Chem.* **1969**, *244*, 4704.
(20) Keil-Dlouha, V.; Zylber, N.; Imhoff, J. M.; Tong, N. T.; Keil, B. *FEBS Lett.* **1971**, *16*, 291.

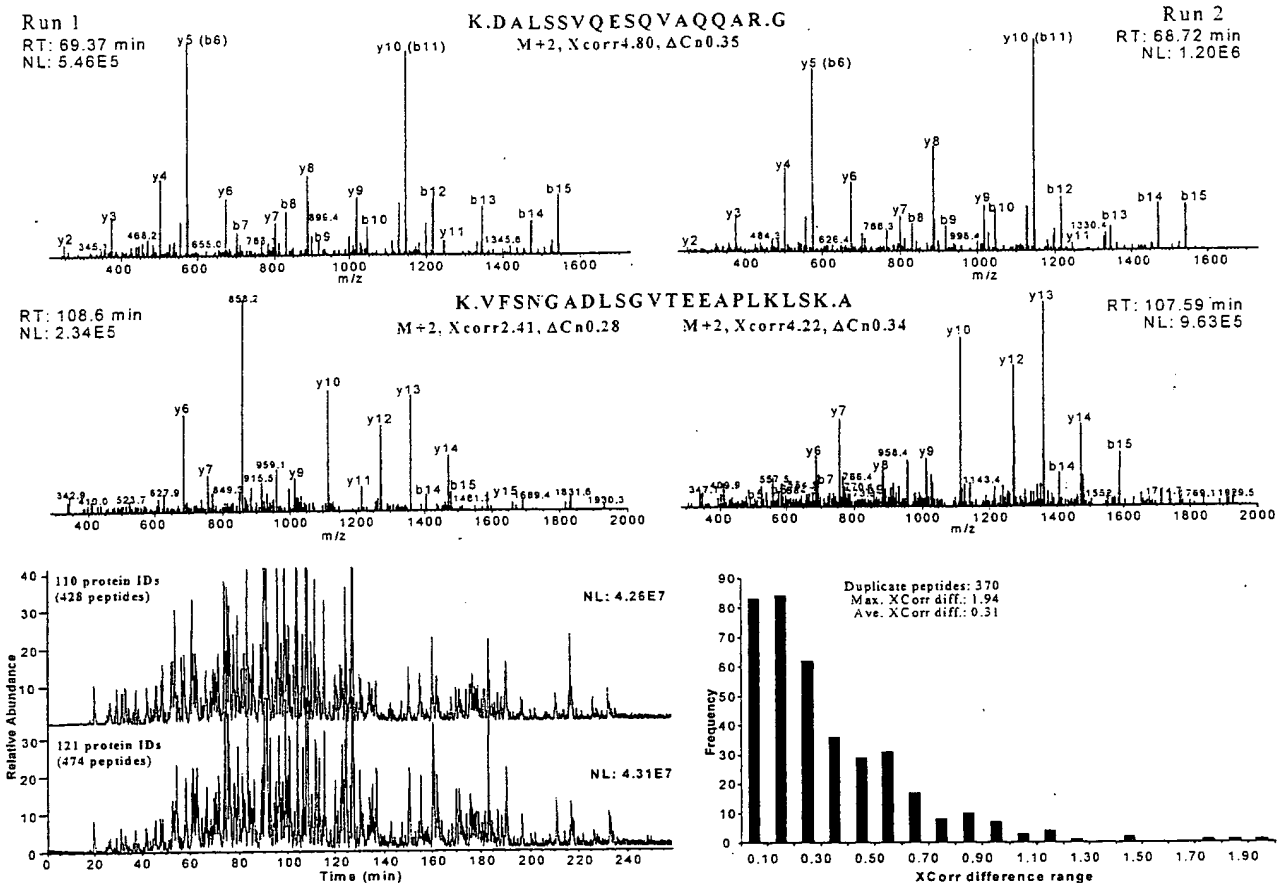

Figure 2. Reproducibility of nanoscale RPLC separations and MS/MS spectra for two successive runs of 0.5 μg of the plasma sample. Reproducible LC separations (bottom left) and MS/MS spectra for some peptides (as shown upper) were obtained, but significant variance of MS/MS spectrum quality for specific peptides were also found (as shown in the middle). The bottom right shows the distribution of X_{corr} variances for peptides assigned from both runs (i.e., duplicate peptides) using the reported peptide identification criteria.[1] Experimental conditions are the same as for Figure 1.

428 different peptides from the 260-min separation. (If not specified, the numbers of peptide and protein identifications given related to evaluation of the methodology were based upon using ΔCn >0.05 as part of the search criteria; see Experimental Section.) This number of IDs is significantly smaller than that usually obtained from microbial global tryptic digests using high-efficiency RPLC/MS/MS[13] and can be ascribed to the large dynamic range of plasma proteins where peptides digested from highly abundant proteins provide significant interference with the detection of peptides from low-abundance proteins.

Examination of the RPLC/MS/MS reproducibility for the human plasma tryptic digest sample revealed that although a similar number of identifications were obtained from repeated RPLC/MS/MS "shotgun" analyses, the peptide and protein IDs were somewhat different. For example, 110 protein IDs were obtained for one run; however, only 90 of the 110 protein IDs were common to another replicate run that yielded 121 protein IDs, i.e., ~80% overlap. The "nonoverlapping" protein IDs were typically from the assignment of single peptides and where MS/MS spectrum quality and the SEQUEST scores varied significantly between the replicate runs. As illustrated in Figure 2, similar chromatograms were obtained for these two runs and similar quality MS/MS spectra were also observed for some peptides (e.g., K.DALSSVQESQVAQQAR.G in Figure 2). However, significant differences in MS/MS spectrum quality and the SEQUEST X_{corr} and ΔCn scores were observed for other peptides (e.g., K.VFSNGADLSGVTEEAPLKLSK.A in Figure 2). This variation in MS/MS spectrum quality (and the resulting SEQUEST) score for some peptides is one reason for the variation in protein IDs from replicate RPLC/MS/MS runs. The observed run-to-run variability in MS/MS spectrum quality for the same peptide is likely associated with the use of the user-selected "exclusion time" after a specific m/z value has been selected for MS/MS, and during which it will not be selected again. Thus, a peptide can be selected for MS/MS well in advance of its peak intensity and where optimum spectrum quality may not be achieved. For the 370 peptide IDs common to these two runs, a maximum difference in X_{corr} of 1.94 was observed, with an average value of 0.31. A much greater difference of 3.0 was found for nonoverlapping peptides, e.g., protein 3D:2pabA, which was identified by a partially tryptic peptide R.YTIAALLSPYSYSTTAVVTNPKE (M + 2, X_{corr} 3.6, ΔCn 0.5) in run 1 but was not observed in run 2 with X_{corr} > 0.6.

A common strategy is to utilize repetitive shotgun MS/MS analyses of the sample to obtain more peptide IDs from one sample and thus increase the overall proteome coverage. Table 1 shows the cumulative number of different peptide and protein IDs generated from the nondepleted plasma sample using RPLC/MS/

Table 1. Peptide/Protein Identification Coverage from High-Efficiency RPLC/MS/MS Runs[a]

run	different peptides (individ run)	different proteins (individ run)	different peptides (cumulative)	different proteins (cumulative)
1	428	110	428	110
2	474	121	532	141
3	499	138	603	176
4	490	138	655	202
5	503	133	708	221
6	427	120	757	240
7	496	133	796	262
8	514	133	840	277
9	509	130	876	290
10	406	105	902	300
11	172	58	906	304
12	482	124	955	315
13	400	113	1015	332
14	383	107	1026	341
15	43	25	1038	342
16	142	53	1059	348
17	125	64	1090	359
18	123	60	1119	372
19	52	34	1124	376
20	45	26	1126	376
21	572	137	1161	389
22	522	123	1331	412
23	517	137	1351	417
24	543	131	1380	433
25	467	121	1408	445
26	532	133	1431	455
27	437	118	1452	464
28	195	59	1452	464

[a] RPLC experimental conditions were described in the Experimental Section. Runs 1−10, repeated RPLC/MS/MS for 0.5 μg of the plasma sample; runs 11−14, repeated RPLC/MS/MS with the sample sizes of 0.15, 1.0, 2.5, and 5.0 μg, respectively; runs 15−20, RPLC/MS/MS of 2.5 μg of the sample with sub-m/z ranges of 400−600, 600−800, 800−1000, 1000−1200, 1200−1400, and 1400−2000, respectively; runs 21−22, 23−24, 25−26, RPLC/MS/MS replicate runs for 1.5 μg of the sample with top 3, 5, and 7 intensity ions for MS/MS; runs 27−28, 15-μm-i.d. nanoscale RPLC/MS/MS system[14] for 0.5 and 0.15 μg of the sample, respectively.

MS. Under the same RPLC/MS/MS conditions (runs 1−10 in Table 1), 300 different protein IDs from 902 different peptide IDs were obtained. On average, an individual run identified 114 different protein IDs with ~82% overlap between two runs. The dependence of peptide and protein IDs on the sample size (runs 11−14) was examined, and it was found that a 0.5−2.5-μg sample size was optimal for peptide identification.

The 400−2000 m/z range was segmented into six ranges (400−600, 600−800, 800−1000, 1000−1200, 1200−1400, 1400−2000 m/z) for potentially enhancing coverage of low-abundance peptides, and the results are shown as runs 15−20 in Table 1. The greatest number of IDs was obtained in the 800−1000 m/z range. The six segmented m/z range runs yielded a combined 455 different peptide assignments covering 128 protein IDs. In comparison, six repeated runs covering the entire m/z range (400−2000 m/z) produced 757 different peptide assignments covering 240 proteins. Thus, for the present study, analyses using segmented m/z ranges were less effective than simple replicate runs for improving coverage of plasma proteins with high-efficiency LC separations. Examination of the protein IDs generated by the segmented m/z range runs revealed that few low-abundance plasma proteins were detected from this m/z-segmentation approach (data not shown).

The selection of either the five or seven most abundant peaks (rather than three) for each MS/MS data-dependent cycle analysis was also investigated. Two RPLC/MS/MS runs (23 and 24) selecting the five most abundant peaks for each MS/MS scan generated 517 and 543 different peptides, respectively, producing 137 and 131 protein IDs, respectively. Of the protein IDs from the five most abundant peaks, there was a 75% overlap (i.e., 102 protein IDs). This protein ID overlap was similar to that obtained when the three most abundant species (runs 21 and 22) were selected. Tandem MS of the seven most abundant peaks for each data-dependent cycle yielded an average of 127 protein IDs with an overlap of ~70% between the two repeated runs (25 and 26). These results indicate that increasing the number of MS/MS analyses for each cycle provided little improvement to protein coverage. New protein IDs were also obtained using a 15-μm nanoLC system[14] (runs 27 and 28) that provides increased sensitivity and decreased sample consumption.

Summing the protein IDs from all 28 runs, high-efficiency RPLC/MS/MS resulted in the detection of 1452 different peptides correlating to a total of 464 protein IDs (using the $\Delta Cn > 0.05$ criteria). Only 64 μg of a nondepleted plasma sample was consumed for the 28 runs. Compared with previously reported results that used conventional 2-D SCXLC/RPLC/MS/MS for a similar depleted human blood serum,[10] this single-dimension high-efficiency RPLC/MS/MS provided similar protein identification coverage with higher confidence identifications (i.e., without considering the chymotryptic or elastic peptides), substantially reduced sample consumption, and a decrease in the total analysis time even for the nondepleted plasma sample.

Human Plasma Proteome Analysis Using High-Efficiency 2-D SCXLC/RPLC/MS/MS. The separation efficiency and reproducibility of single-dimension high-efficiency RPLC has been previously evaluated,[12,13] and a chromatographic peak capacity of ~1000 can be supplied by single-dimension RPLC. This chromatographic peak capacity can be extended by > 10-fold through combination of the high-efficiency RPLC with orthogonal SCXLC to form a SCX/RP 2-D LC separation. Figure 3 shows a SCXLC/UV chromatogram for separation of a 150-μg plasma sample and RPLC/MS/MS analyses of the resulting 15 fractions. The SCXLC was completed using a highly hydrophilic (2-sulfoethyl aspartamide) stationary phase,[21,22] which is broadly effective for eluting peptides without addition of organic solvents to the mobile phases and that would subsequently affect the remaining steps of an on-line SCX/RP 2-D LC separation. A few chromatographic peaks (monitored by UV at 215 nm) were observed and collected into the first three fractions when the sample-loaded SCX column was washed with mobile phase A (possibly the result of some slight sample overloading). Each SCXLC fraction contained more than one apparent chromatographic UV peak, and the combination of 15 RPLC/MS/MS runs from the SCXLC fractions led to a conservative estimate of $>10^4$ for the total SCX/RP 2-D LC separation peak capacity. Highly nonuniform distributions of protein IDs were found for the multidimensional SCXLC/RPLC/MS/MS analyses (the number of protein IDs for each fraction is given in parentheses in their LC/MS/MS base peak chromatograms of Figure 3), with the greatest number of protein IDs (208)

(21) Alpert, A. J.; Andrews, P. C. *J. Chromatogr.* **1988**, *443*, 85−96.
(22) Crimmins, D. L. *Anal. Chim. Acta* **1997**, *352*, 21−30.

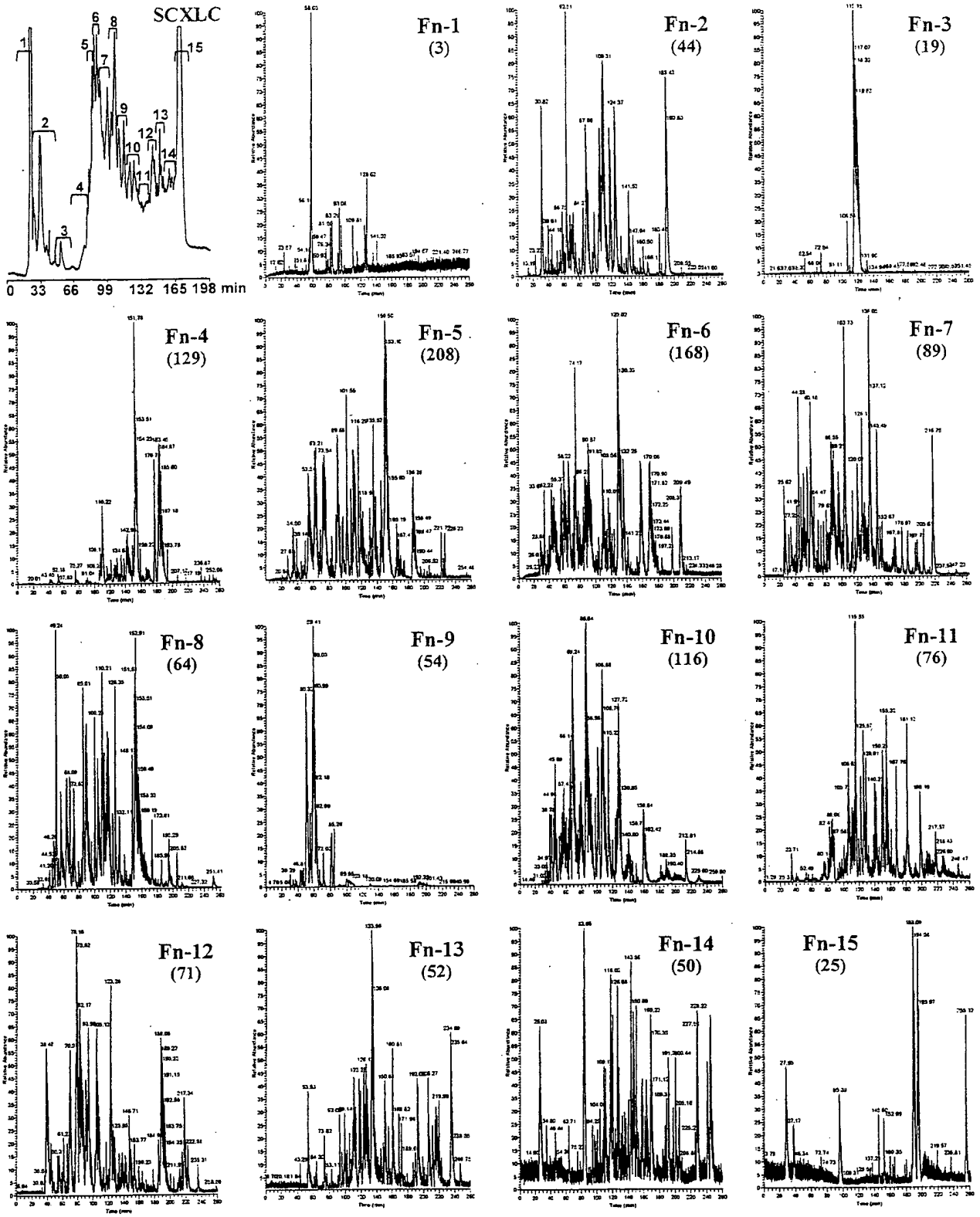

Figure 3. High-efficiency SCXLC/RPLC/MS/MS of nondepleted human plasma tryptic digest. The nondepleted human plasma sample (150 μg) was separated into 15 fractions followed by RPLC/MS/MS runs of the total or a portion of each fraction (based upon the fraction peak intensities). The number of protein IDs for the corresponding fraction are given in parentheses.

obtained from fraction 5 and the fewest (3) from fraction 1. (The chromatographic intensity scale was based upon the most abundant peak; thus there is no strong correlation between the obvious complexity of the chromatograms as shown and the number of

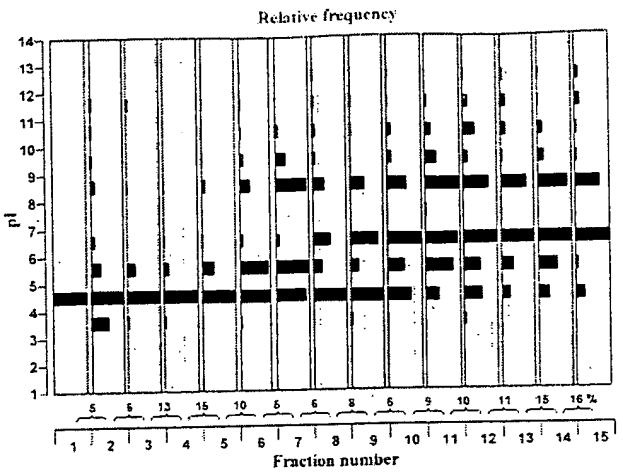

Figure 4. Calculated p*I* distribution for identified peptides from the SCXLC fractions (see Figure 3). The number across two adjacent fractions gives the percentage overlap for peptides identified between the two adjacent fractions.

protein IDs.) The 15 SCXLC fractions provided 593 protein IDs (using ΔCn >0.05 criteria), which is 28% greater than that achieved using the single-dimension RPLC/MS/MS approach described above (464 protein IDs).

The distribution of isoelectric points (p*I*) of peptides identified from the 15 fractions is shown in Figure 4. The peptide overlap between two adjacent fractions averaged ~10% (ranging from 5 to 16%), demonstrating the effectiveness of the SCXLC sample fractionation. Generally, an increase of average peptide p*I*s with increasing fraction number was observed; however, there was no strong relationship between the p*I*s of individual peptides and their SCXLC fraction. Few plasma tryptic peptides had p*I* values between 7 and 8; peptides having widely dispersed p*I* values (e.g., 3–13) were found in single fractions (e.g., fraction 12); and some peptides having calculated p*I* values between 3 and 4 were identified in all 15 SCXLC fractions. These observations reveal that, in addition to peptide charge, other properties (e.g., hydrophilicity and solubility) also contribute to SCXLC retention. While the technique of capillary isoelectric focusing provides a more direct p*I* dependence fpr peptide separation.[4,23] SCXLC provides a substantially larger p*I* range for eluting peptides (e.g., p*I* >12 or p*I* <4) in a single separation run. The relationship between peptide p*I* and SCXLC elution can be potentially illuminated using the artificial neural network approach.[17]

The SCXLC fractions obtained from a total of 150 μg of plasma protein were sufficient to enable multiple RPLC/MS/MS runs for each SCXLC fraction (generally 0.5–2.5 μg, far larger than that used for a single 30-μm-i.d. column run). In Figure 5 the number "*XXX*/*n*" marked in parentheses next to the SCXLC chromatogram shows the number of protein IDs from RPLC/MS/MS runs for the fraction. The fact that a single RPLC/MS/MS analysis does not provide the identification of all potentially amenable peptides in the sample is clearly demonstrated by the insets for fractions 10 and 13. For example, a single RPLC/MS/MS run for fraction 10 obtained a range of 95–116 protein IDs per run; however, a combined total of 213 different protein IDs were identified from five repeated RPLC/MS/MS runs (overlaps between two runs were ~70%). This general phenomenon was observed for all of the fractions investigated (e.g., compare the numbers given in the parentheses in Figures 3 and 5), independent of the number of protein IDs they contained. For SCXLC fractions with small protein content (indicated by their SCXLC–UV peak intensity), the fraction solution was concentrated to 10-μL volume for a single RPLC/MS/MS run, as demonstrated for fraction 3 in Figure 5.

Summing protein IDs identified from each run (39 runs total), this SCXLC/RPLC/MS/MS approach identified 881 protein IDs (using the ΔCn >0.05 criteria), a ~49% increase compared with single run (593 protein IDs) of each SCXLC fraction. We repeated another cycle of SCXLC/RPLC/MS/MS and 198 new protein IDs (i.e., ~20% increase in the number of protein IDs) were obtained, leading to the total identification of 1076 protein IDs from the SCXLC/RPLC/MS/MS. These protein IDs covered only 62% of those obtained from the single-dimension RPLC/MS/MS. The limited number RPLC/MS/MS runs for each SCXLC fraction (a total of 77 RPLC/MS/MS runs for two cycles of SCXLC with 15 fractions, compared to 28 RPLC/MS/MS runs for the single whole sample) is one contributor for this relatively low coverage. For example, the indicated coverage is increased to 78% if the protein IDs from the first six (i.e., averaged number of runs for each SCXLC fraction) RPLC/MS/MS runs with the whole sample are counted. The other factor that influences coverage includes the possible sample losses from the SCXLC process. Therefore, replicate RPLC/MS/MS runs are complementary to SCXLC/RPLC/MS/MS for minimal MS/MS detection of missing peptides/proteins, and its complementary capability is determined by the RPLC separation power (e.g., separation peak capacity).

Description of Human Plasma Proteins Identified. Combining protein IDs identified from the high-efficiency RPLC/MS/MS and high-efficiency SCXLC/RPLC/MS/MS, a total of 1348 protein IDs were obtained from assignments of 3319 different peptides by 105 RPLC/MS/MS runs from ~365 μg of the nondepleted human plasma sample. This total is based upon previously published SEQUEST search criteria[1] varying in only the ΔCn, which was set at >0.05. Also, these identifying protein IDs were examined manually in the database for possible redundancies, and numerous instances were found where the same protein was contained in multiple database protein IDs. This redundancy was removed, reducing the 1348 protein IDs down to a set of 1235 proteins (Table 3). Furthermore, to obtain a higher confidence list of detected proteins, peptides were further filtered to eliminate peptide identifications that had a predicted normalized LC elution time[17] with a difference greater than ±10% of the measured theoretical value (see Experimental Section), significantly increasing the overall confidence of the data set. Such analyses have been previously shown to successfully remove lower confidence peptide identifications from human data sets.[24] Table 3 shows the total number of proteins based upon use of either a ΔCn >0.05 or a >0.1 value, as well as a comparison with two additional SEQUEST filter criteria that have been applied previously.[10,16,36] In addition to the NET constraint, a more stringent definition of "partial tryptic" was added to all criteria to further reduce false identifications of peptides from incomplete digestion

(23) Shen, Y.; Berger, S. J.; Anderson, G. A.; Smith, R. D. *Anal. Chem.* 2000, *72*, 2154–2159.

(24) Jacobs, J. M.; Mottaz, H. M.; Yu, L.; Anderson, D. J.; Moore, R. J.; Chen, W. U.; Auberry, K. J.; Strittmatter, E. F.; Monroe, M. E.; Thrall, B. D.; Camp, D. G.; Smith, R. D. *J. Proteome Res.*, in press.

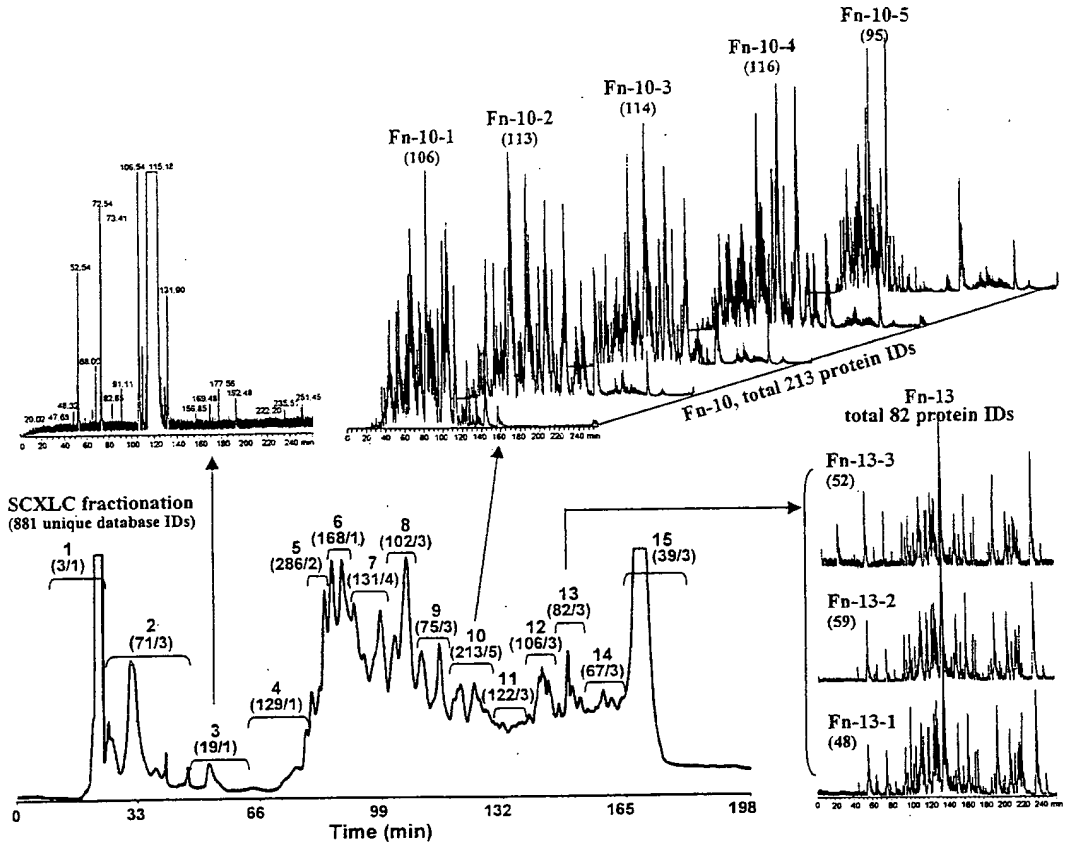

Figure 5. Replicate RPLC/MS/MS analyses of nondepleted human plasma tryptic digest using high-efficiency SCXLC/RPLC/MS/MS. Replicate runs using the remaining samples from each fraction, after achieving results shown in Figure 3, were analyzed using the same RPLC/MS/MS conditions. The values given in parentheses represent the number of protein IDs, and *XXX*/*n* represents *XXX* IDs obtained by *n* runs. The SCXLC, RPLC, and ion trap MS/MS experimental conditions are described in the Experimental Section.

(see Experimental Section). Based upon this analysis, a total of 1061 proteins are reported. The application of alternative and still more conservative criteria reduces the set further (to 800 for the combination of the NET constraint and $\Delta Cn > 0.1$). A listing of all 1235 proteins and the corresponding peptides with their SEQUEST scores are given in the Supporting Information.

The incorporation of a lower ΔCn was explored because of concerns that by selecting a stricter ΔCn cutoff of >0.1 many identifications would be excluded during the data analysis (but inevitably with the possibility of contributing false identifications). However, we also implemented the additional criteria based upon peptide LC NET and more stringent "partial tryptic rules". A comparison of the effects of applying the two different ΔCn cutoff values and also the NET constraint is shown in Table 3. We observed that a large number of proteins removed using the NET criterion were also identified using only one peptide, consistent with the expectation that these identifications were less confident.

Figure 6 shows a distribution of human plasma proteins identified from the nondepleted sample using both RPLC/MS/MS and SCXLC/RPLC/MS/MS. Cellular (leakage) proteins correspond to those proteins that would normally be found within a cell and are presumably present in the plasma due to leakage. The categorization of "classic" plasma proteins corresponds to previously characterized proteins that are specifically localized for activity in plasma (i.e., human serum albumin, complement components, and apolipoproteins).[8] Immunoglobulin identifications accounted for the largest percentage (38%). Approximately 11% of cellular proteins were determined to be secreted or extracellular, with 3% identified as cytokines or cytokine-related proteins. Approximately 8% of such proteins are known to be extracellular, but not necessarily localized in the plasma. The data show that even in the presence of the high-abundance serum albumin and immunoglobulin proteins, 642 additional proteins (using $\Delta Cn > 0.05$ criteria and not including immunoglobulins) were identified. This ability to detect a wide range of proteins for nondepleted plasma samples is significant, since any depletion procedure potentially results in selective protein losses.

Human Plasma Proteome Analysis Dynamic Range. In addition to the separation efficiency and analysis sensitivity, the $\sim 10^2$ sample preconcentration after SCXLC fractionation contributed significantly to the extension of ion trap MS/MS protein identification dynamic range. The present studies were also enabled by the increased ruggedness provided by use of on-line microSPE sample manipulation in conjunction with the nanoscale RPLC.

Using the published data available for human plasma proteins, the dynamic range represented by the set of identified proteins was examined. Table 2 gives selected proteins identified at various concentration levels in human plasma. Human serum albumin, the most abundant protein in human plasma is present at 35−55 mg/mL of plasma.[25] The lowest abundance protein identified using the high-efficiency RPLC/MS/MS was 5 μg/mL of plasma

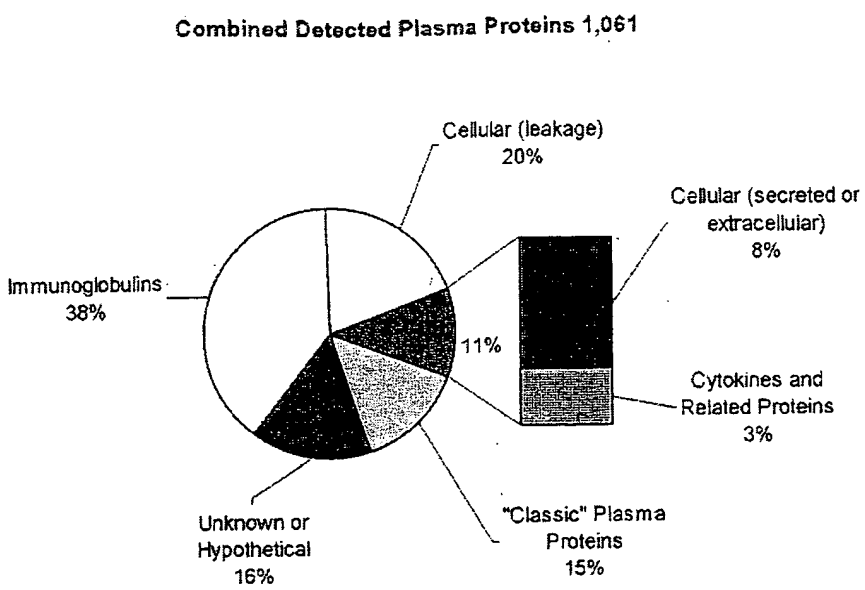

Figure 6. Categories of detected proteins from the human plasma sample studied.

Table 2. Examples of Proteins Identified from a Nondepleted Human Plasma Sample at Various Abundance Levels[a]

	High-Efficiency RPLC/MS/MS
mg level	35−50 mg of albumin 3.8−7.8 mg of haptoglobin, 2−4.5 mg of fibrinogen 2−4 mg of transferrin, α-1-anti-trypsin 1.6−3.8 mg of α-2-macroglobulin
μg level	200−700 μg of inter-α-trypsin inhibitor, 300−600 μg of α-1-anti-chymotrypsin 300 μg of fibronectin, 40−150 μg of complement C5, 70−90 μg of complement C8 60−80 μg of transcortin, 50 μg of complement C1r 10−30 μg of complement C2, 10−20 μg of coagulation factor XIII 5 μg of pigment epithelial-derived factor (PEDF)
	High-Efficiency SCXLC/RPLC/MS/MS
ng level	~200 ng of matrix metalloproteinase-2 (MMP-2) 40−100 ng of T-lymphocyte activation antigen (CD80) ~80 ng of hepatocyte growth factor activator (HGFA) 10−30 ng of macrophage stimulatory protein (MSP) ~1 ng of human megakaryote stimulating factor (MSF) ~1 ng of interleukin-1 receptor (IL-1 R)
pg level	77 pg of interleukin-12 β chain (IL-12 p40) ~10−30 pg of fibroblast growth factor-12 (FGF-12)

[a] For high-efficiency SCXLC/RPLC/MS/MS, only low-abundance (i.e., < ng level) proteins identified are listed. The concentrations (in unit mL) are based upon either product literature from R&D Systems Quantikine Immunosassay Kits or the literature.[25−33] Concentrations of these proteins are known to vary from sample to sample so the concentration shown here is an approximate level at which such proteins are detected in plasma. All listed proteins have ΔCn values of >0.1, and passed the NET criteria.

Table 3. Number of Peptides and Proteins Identified Using Different Criteria

criteria	ref 1[a] ΔCn >0.1	ref 1[a] ΔCn >0.05	ref 16[b]	refs 10 and 36[c]
peptides	2912	3268	3308	3935
proteins	880	1235	1193	1682
proteins removed by peptide NET constraint[d]	80	174	206	389
total proteins	800	1,061	987	1293

[a] From ref 1. Criteria: $X_{corr} \geq 1.9$ for 1+ and full tryptic cleavage, $X_{corr} \geq 2.2$ for 2+ and full or partial tryptic cleavage, $X_{corr} \geq 3$ for 2+ for all peptides, and $X_{corr} > 3.75$ for 3+ and full or partial tryptic cleavage. [b] From ref 16. Criteria are that all peptides at least "half" tryptic and ΔCn ≥0.08 plus the following: $X_{corr} \geq 1.8$ for +1, 2.5 for +2, and 3.5 for +3. [c] From refs 10 and 36. Criteria are as follows: $X_{corr} \geq 1.9$ for 1+ and full tryptic, $X_{corr} \geq 2.1$ for 1+ and chymotryptic or elastic, $X_{corr} \geq 2.2$ for +1 and partially tryptic, chymotryptic, or elastic, $X_{corr} \geq 2.2$ for 2+ and full tryptic, $X_{corr} \geq 2.4$ for 2+ and partially tryptic, chymotryptic or elastic, $X_{corr} \geq 3.0$ for 2+ for all peptides, and $X_{corr} \geq 3.75$ for 3+ and tryptic, chymotryptic, or elastic. [d] Proteins removed when NET agreement for corresponding peptides is not within 10% of predicted value (see Experimental Section). All: "partially tryptic" refers to peptides meeting a stricter definition (see Experimental Section).

pigment epithelial-derived factor[26] that was identified by assignment of the tryptic peptide R.DTDTGALLFIGK.I with M + 2, X_{corr} of 2.5, and ΔCn of 0.2, thus indicating a protein identification dynamic range of ~4 orders of magnitude in protein abundance.

This dynamic range is in good agreement with our previous work using high-efficiency RPLC/MS/MS,[14] and equivalent to that obtained using conventional 2-D SCXLC/RPLC/MS/MS.[2] The lowest abundance protein identified in the present work was fibroblast growth factor 12 (FGF-12) that exists at ~10−30 pg/mL of plasma.[27,28] The identification of this protein was based upon assignment of its partially tryptic peptide V.AMNGEGYLYSSD-VFTPECK.F (see MS/MS spectrum shown in Figure 7) with M + 2, X_{corr} of 2.6, and ΔCn of 0.107 according to the peptide assignment criteria.[1,16]

(25) Higher abundant plasma protein values were taken from: Craig, W. Y.; Ledue, T. B.; Ritchie, R. F. *Plasma Proteins: Clinical Utility and Interpretation.* Provided online by the Foundation for Blood Research (FBR) (http://www.fbr.org/). And: Putnam, F. W., Ed. *The plasma proteins: structure, function, and genetic control.* 2nd ed.; Academic Press: New York, 1975.

(26) Peterson, S. V.; Valnickova, Z.; Enghild J. J. *Biochem. J.* **2003**, *374*, 199−206.

(27) Based upon values of similar FGF growth factors provided by R&D Systems literature; Plasma values for IL-12, MSP, MMP-2, and CD80 were obtained from product literature provided by R&D Systems Quantikine immunoassay kits.

(28) Smallwood, P. M.; Munoz-Sanjuan, I.; Tong, P.; Macke, J. P.; Hendry, S. H. C.; Gilbert, D. J.; Copeland, N. G.; Jenkins, N. A.; Nathans, J. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 9850−9857.

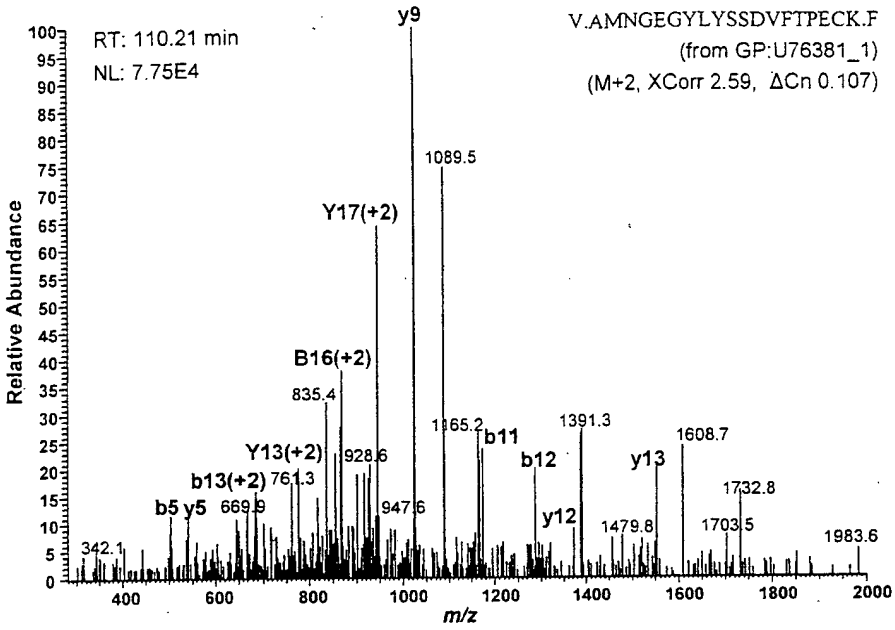

Figure 7. Identification of FGF-12 at a content of ~10−30 pg/mL of plasma from fraction 9 of the SCXLC/RPLC/MS/MS analysis (see Figure 5).

The successful identification of ~10 pg/mL of plasma-level proteins from 150 μg of the nondepleted plasma sample (used in the SCXLC/RPLC/MS/MS) allows estimation of the detection sensitivity of ~10 amol, assuming an average protein MW of ~20 000 and the nondepleted plasma protein content of 75 mg/mL. The RPLC/MS/MS system used in this study provides a peptide detection sensitivity of ~5 amol based upon previous investigations using 15-μm-i.d. column RPLC/ion trap MS/MS[14] and the expected relationship between the MS sensitivity and LC column inner diameter.[12] Even though almost all of the identified peptides corresponding to proteins in Table 2 easily surpassed our most stringent filter criteria, detection of low-abundance species supporting a $>10^9$ dynamic range relies upon proteins identified based upon only one peptide. A conservatively estimated dynamic range of $\sim 0.5 \times 10^8$ is based upon the high confidence identifications of several fully tryptic peptides from the protein MSF,[29] that include K.SEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPR.V (M + 3, X_{corr} 5.09, ΔCn 0.4), R.GLPNVVTSAISLPNIR.K (M + 2, X_{corr} 5.0, ΔCn 0.4), R.AIGPSQTHTIR.I (M + 2, X_{corr} 2.8, ΔCn 0.3), and K.DQYYNIDVPSR.T (M + 2, X_{corr} 2.4, ΔCn 0.1). The results presented here demonstrate that implementation of high-efficiency separations prior to MS detection can facilitate detection of trace components that have a dynamic range of >8 orders of magnitude in plasma concentration.

CONCLUSIONS

The present results demonstrate that high-efficiency SCX/RP 2-D LC separations (having combined peak capacities of $>10^4$) greatly extend the protein identification dynamic range of conventional ion trap MS/MS to greater that 8 orders of magnitude in protein relative abundance. This dynamic range enabled the identification of >800 human proteins from ~3000 different peptides without the depletion of abundant plasma proteins. The dynamic range extension enabled the identification of picogram per milliliter of plasma proteins such as cytokines and related proteins in the presence of human serum albumin. The use of 30-μm-i.d. nanoscale RPLC columns limited the total sample consumption to 365 μg of a nondepleted human plasma sample over a series of 105 RPLC/MS/MS runs. The high-efficiency (i.e., peak capacity of $\sim 10^3$) single-dimension RPLC/MS/MS analyses provided a dynamic range of 4−5 orders of magnitude, producing 464 protein IDs from assignment of 1452 peptides. As previously reported, replicate RPLC/MS/MS runs are advantageous for improving plasma proteome coverage from analysis of the plasma peptide SCXLC fractions, at least partially due to the run-to-run variations in MS/MS spectrum quality, especially for low-abundance proteins. The incomplete overlap of proteins detected from RPLC/MS/MS and SCXLC/RPLC/MS/MS suggests the utility of using both approaches for optimizing coverage.

Planned efforts will examine the expansion of coverage obtained using alternative fractionation methods. The combination of these LC/MS/MS analyses will provide the basis for a comprehensive peptide mass and time tag lookup table (i.e., the peptide mass vs. its RPLC retention time from RPLC/MS/MS) for human plasma and the foundation for LC/Fourier transform ion cyclotron resonance MS high-throughput analyses.[13,34] This approach is expected to provide sensitive (e.g., based on nanogram-level samples),[14] wide dynamic range, and quantitative (i.e.,

(29) Tayrien, G.; Rosenberg, R. D. *J. Biol. Chem.* **1987**, *262*, 3262−3268.
(30) Gerhardt, W.; Ljungdahl, L. *Clin. Chim. Acta* **1998**, *272*, 47−57.
(31) Luo, J. C.; Neugut, A. I.; Garbowski, G.; Forde, K. A.; Treat, M.; Smith, S.; Carney, W. P.; Brandt-Rauf, P. W. *Cancer Lett.* **1995**, *91*, 235−240.
(32) Miyazawa, K.; Shimomura, T.; Kitamura, A.; Kondo, J.; Morimoto, Y.; Kitamura, N. *J. Biol. Chem.* **1993**, *268*, 10024−10028.
(33) Giri, J. G.; Wells, J.; Dower, S. K.; McCall, C. E.; Guzman, R. N.; Slack, J.; Bird, T. A.; Shanebeck, K.; Grabstein, K. H *J. Immunol.* **1994**, *153*, 5802−5809.
(34) Shen, Y.; Tolić, N.; Zhao, R.; Paša-Tolić, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 3011−3021.

using isotopically labeled peptides)[34,35] analyses with sufficient throughput to enable the characterization of sufficient numbers of plasma samples to identify new protein biomarkers that are associated with trauma or that are diagnostic of individuals either predisposed to disease states, or the disease states themselves.

ACKNOWLEDGMENT

We thank the National Institute of General Medical Sciences (NIGMS, Large Scale Collaborative Research Grants U54 GM-62119-02), the NIH National Center for Research Resources (RR018522), and the Environmental Molecular Sciences Laboratory at PNNL for the support of portions of this research. Pacific Northwest National Laboratory is operated by the Battelle Memorial Institute for the U.S. Department of Energy through Contract DE-AC06-76RLO 1830.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review July 28, 2003. Accepted November 26, 2003.

AC034869M

(35) Smith, R. D.; Paša-Tolić, L.; Lipton, M. S.; Jensen, P. K.; Anderson, G. A.; Shen, Y.; Conrads, T. P.; Udseth, H.; Harkewicz, R.; Belov, M. E.; Masselon, C.; Veenstra, T. D. *Electrophoresis* **2001**, *22*, 1652−1668.

(36) Tirumalai, R. S.; Chan, K. C.; Prieto, D. A.; Issaq, H. J.; Conrads, T. P.; Veenstra, T. D. *Mol. Cell. Proteomics*, in press.

RESEARCH ARTICLE

Large-scale analysis of the yeast proteome by multidimensional protein identification technology

Michael P. Washburn[1†], Dirk Wolters[1†], and John R. Yates III[1,2*]

We describe a largely unbiased method for rapid and large-scale proteome analysis by multidimensional liquid chromatography, tandem mass spectrometry, and database searching by the SEQUEST algorithm, named multidimensional protein identification technology (MudPIT). MudPIT was applied to the proteome of the *Saccharomyces cerevisiae* strain BJ5460 grown to mid-log phase and yielded the largest proteome analysis to date. A total of 1,484 proteins were detected and identified. Categorization of these hits demonstrated the ability of this technology to detect and identify proteins rarely seen in proteome analysis, including low-abundance proteins like transcription factors and protein kinases. Furthermore, we identified 131 proteins with three or more predicted transmembrane domains, which allowed us to map the soluble domains of many of the integral membrane proteins. MudPIT is useful for proteome analysis and may be specifically applied to integral membrane proteins to obtain detailed biochemical information on this unwieldy class of proteins.

Modern biologists can now observe quantitative changes in the expression levels of thousands of messenger RNA (mRNA) transcripts to determine the effects of a wide variety of perturbations to a cell[1]. However, there exists conflicting evidence regarding the correlation between mRNA and protein abundance levels[2–5]. Recent mathematical modeling studies have demonstrated the need to know both the mRNA and protein expression levels of genes in order to describe a gene network[6,7]. The need to complement mRNA expression analysis has resulted in the emergence of the field of proteomics to directly analyze protein expression levels from an organism.

The analysis of a proteome requires the resolution of the proteins in a sample followed by the identification of the resolved proteins. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) followed by mass spectrometry (MS) is the most widely used method of protein resolution and identification[8–10]. In 2D-PAGE, proteins are separated in one dimension by isoelectric point (pI) and in the other dimension by molecular weight (MW). High-throughput analysis of proteomes remains challenging because the individual extraction, digestion, and analysis of each spot from 2D-PAGE is a tedious and time-consuming process. As a result, the largest 2D-PAGE-based proteomic study to date identified 502 unique proteins for the *Haemophilus influenzae* proteome[11]. Portions of proteomes such as proteins with extremes in pI and molecular weight[12,13], low-abundance proteins[14–16], and membrane-associated or bound proteins[17,18] are rarely seen in a 2D-PAGE study. While efforts to alleviate the current shortcomings in 2D-PAGE continue, we are exploring non-gel-based chromatography systems to resolve and identify thousands of proteins from a biological sample[19–21].

Like 2D-PAGE, an alternative two-dimensional separation system must subject proteins or peptides to two independent separation methods and maintain the separation of two components after they have been resolved in one step[22]. A variety of efforts are underway to utilize multidimensional chromatography coupled with mass spectrometry to characterize proteomes[23]. Link *et al.* developed an online method coupling two-dimensional liquid chromatography (LC) to tandem mass spectrometry (MS/MS) (Fig. 1)[19]. In this method a pulled microcapillary column is packed with two independent chromatography phases[19]. Once a complex peptide mixture was loaded onto the system, no additional sample handling was required because the peptides eluted directly off the column and into the mass spectrometer (Fig. 1)[19]. After optimizing this system, we carried out the largest number of protein identifications in any proteome to date. By simultaneously resolving peptides and identifying their respective proteins, the system separated and identified 1,484 proteins from the *S. cerevisiae* proteome. Because the system is largely unbiased, proteins from all subcellular portions of the cell with extremes in pI, MW, abundance, and hydrophobicity were identified.

Results

The MudPIT method described is reproducible on the levels of both the chromatography and the final protein list (data not shown). Chromatographic reproducibility is described as the identification of the same peptide at the same point in the chromatography in two or more separate analyses. The results reported in this paper are from representative runs of the three separate fractions. After combining the MS/MS data generated from all three different samples, we were able to assign 5,540 peptides to MS spectra leading to the identification of 1,484 proteins from the *S. cerevisiae* proteome. A complete list of the proteins and peptides identified is available as Supplementary Table 1 in the Web Extras page of *Nature Biotechnology* Online. Each of the three preparations (soluble fraction, lightly washed insoluble fraction, and heavily washed insoluble fraction) provided unique hits to the final data set. The proteins identified in the AUTOQUEST output were further analyzed using the MIPS *S. cerevisiae* catalogs[24]. This analysis revealed that (1) our results provide a representative sampling of the yeast proteome, and (2) our MudPIT method is largely unbiased, meaning that low-abundance proteins, proteins with extremes in pI and MW, and integral membrane proteins were identified with the same sensitivity as any other protein.

[1]*Syngenta Agricultural Discovery Institute, 3115 Merryfield Row, Suite 100, San Diego, CA 92121.* [2]*Department of Cell Biology SR11, 10550 North Torrey Pines Road, The Scripps Research Institute, La Jolla, CA 92037.* **Corresponding author (jyates@scripps.edu).* [†]*These authors contributed equally to this work.*

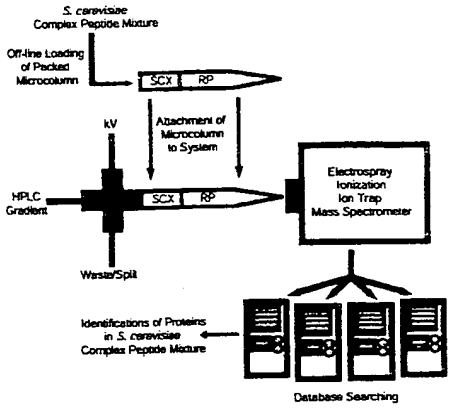

Figure 1. Multidimensional protein identification technology (MudPIT). Based on the method of Link *et al.*[19], complex peptide mixtures from different fractions of a *S. cerevisiae* whole-cell lysate were loaded separately onto a biphasic microcapillary column packed with strong cation exchange (SCX) and reverse-phase (RP) materials. After loading the complex peptide mixture into the microcapillary column, the column was inserted into the instrumental setup. Xcalibur software, HPLC, and mass spectrometer were controlled simultaneously by means of the user interface of the mass spectrometer. Peptides directly eluted into the tandem mass spectrometer because a voltage (kV) supply is directly interfaced with the microcapillary column. As described in the Experimental Protocol, peptides were first displaced from the SCX to the RP by a salt gradient and eluted off the RP into the MS/MS. In an iterative process, the microcolumn was re-equilibrated and an additional salt step of higher concentration displaced peptides from the SCX to the RP. Peptides were again eluted by an RP gradient into the MS/MS, and the process was repeated. The tandem mass spectra generated were correlated to theoretical mass spectra generated from protein or DNA databases by the SEQUEST algorithm[21].

Representative sampling of the yeast proteome. The subcellular localization catalogs from MIPS (ref. 24) allowed us to determine the similarities and differences among the three fractions (Table 1). Even though in several cases the overall numbers of proteins identified from a cellular compartment appear similar between any two samples, unique identifications were found in every sample. For example, the majority of the unique hits from the soluble fraction were proteins localized to the cytoplasm and the nuclei of *S. cerevisiae* including the transcription factor SNF5 (Codon Adaptation Index[25] (CAI) = 0.12)[26,27] and the superoxide dismutase chaperone LYS7 (CAI = 0.16)[28].

The two insoluble fractions provided greater detections and identifications of organelle proteins (Table 1). The heavily washed insoluble fraction had more hits than any other sample localized to the nucleus, mitochondria, endoplasmic reticulum, plasma membrane, and Golgi (Table 1). There were unique hits found in both the heavily washed insoluble fraction and partially washed insoluble fraction. For example, the majority of the hits to the vacuole were identified in the lightly washed insoluble fraction including the H^+-ATPase domains VMA4 (CAI = 0.27) and VMA5 (CAI = 0.24)[29].

Using the MIPS catalogs we determined that every major functional category and protein class were represented in our data (data not shown). Of the major protein classes rarely seen on 2D-PAGE, we detected and identified 32 protein kinases including the MAP kinase signal transduction pathway kinases STE7 (CAI = 0.12), STE11 (CAI = 0.15), STE20 (CAI = 0.16), and FUS3 (CAI = 0.12)[30,31]. Furthermore, we detected and identified 45 transcription factors including members of the SWI–SNF complex SNF5 (CAI = 0.12), SWI4 (CAI = 0.15), and SWI6 (CAI = 0.14)[26,27].

Of the 6,216 open reading frames in the yeast genome, 83% have CAI values between 0 and 0.20, that is, are predicted to be present at low levels (Fig. 2A). Previous proteomics studies in yeast have identified few proteins with CAIs <0.2 (refs. 4,5,32). Efforts are underway to overcome these shortcomings of 2D-PAGE, but recent evidence suggests that 2D-PAGE alone is incapable of detecting low-abundance proteins[16]. Any large-scale proteomic analysis of *S. cerevisiae* must identify proteins in this CAI range. As seen in Figure 2B, the data from our study yield a representative sample of the yeast proteome with 791 or 53.3% of the proteins identified having a CAI of <0.2. A total of 1,347 peptides were detected from the 791 proteins identified with a CAI of <0.2, an average of 1.7 peptides per protein. The number of peptides per protein increases with increasing CAI (Fig. 2C). Because CAI is considered a predictor of protein abundance[4], the most abundant proteins are the easiest to detect in any sample resulting in more peptide identifications from abundant proteins than low-abundance proteins.

Extremes of the *S. cerevisiae* proteome are well represented in our data. Because a peptide mixture is generated before the chromatography, the method should be independent of pI and MW of proteins. In two of the studies for which MW and pI were reported for the proteins identified, no protein with a MW >180 kDa or pI >10 was detected and identified[5,32]. Proteins with both acidic and basic pIs are represented in our data set. Twelve proteins with pIs <4.3 were identified, with the lowest being RPP1A (YDL081C), which has a pI of 3.82 (data not shown). Twenty-nine proteins with pIs >11 were identified, with the most basic protein identified being RPL39 (YJL189W), which has a pI of 12.55 (data not shown). In addition, proteins with MWs <10,000 and >190,000 Da are represented. For example, 24 out of 77 possible proteins with a MW in excess of 190 kDa were identified, the largest being YLR106C (CAI = 0.17) with a MW of 558,942 Da, from which four unique peptides were identified.

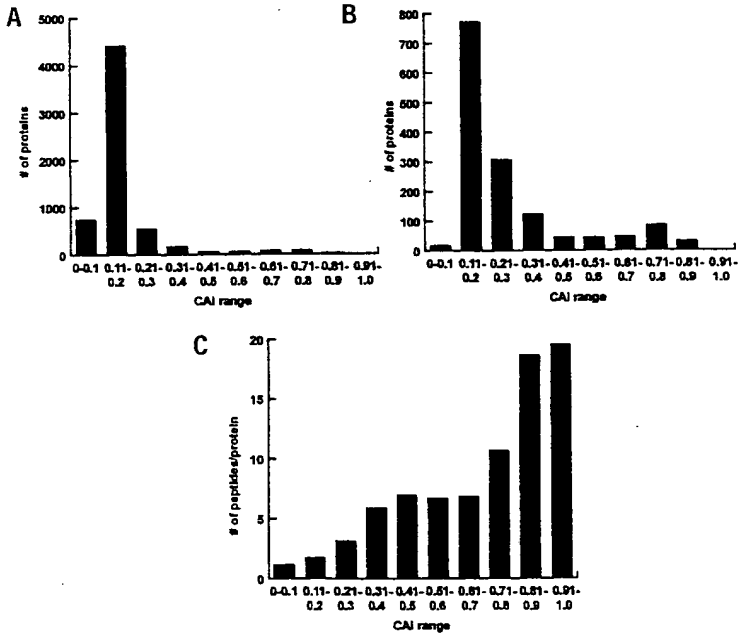

Figure 2. Codon adaptation index (CAI) distribution of the identified *S. cerevisiae* proteome and the predicted *S. cerevisiae* genome. (A) CAI distribution of the proteins predicted in the *S. cerevisiae* genome. (B) Compare this to the distribution of the proteins identified in this study over CAI ranges. In both cases, the largest protein region is found between the CAI range of 0.11 and 0.2. (C) The average number of peptides identified for each protein in a particular CAI range was determined and plotted against CAI ranges.

Table 1. Known subcellular localization of proteins identified in *S. cerevisiae* fractions[a]

Subcellular compartment	Soluble fraction[b]	Lightly washed insoluble fraction[b]	Heavily washed insoluble fraction[b]
Cell wall	2	1	1
Plasma membrane	5	18	35
Cytoplasm	286	264	274
Cytoskeleton	11	20	22
Endoplasmic reticulum	12	36	42
Golgi	3	10	16
Transport vesicles	4	14	16
Nucleus	67	122	151
Mitochondria	43	87	83
Peroxisome	2	3	3
Endosome	1	1	2
Vacuole	5	10	6
Microsomes	0	0	1
Lipid particles	0	2	3

[a]Subcellular localizations obtained from the *S. cerevisiae* subcellular localization catalog at the Munich Information Center for Protein Sequences website [24].
[b]Proteins identified in individual runs were analyzed for their subcellular localization. The subcellular localization of many of the proteins detected and identified is unknown. Therefore, not all of the proteins detected and identified are represented in this table.

Detection and identification of integral and peripheral membrane proteins. By analyzing our data set against the peripheral membrane proteins contained in the Yeast Proteome Database[33], we detected and identified 72 out of 231 possible peripheral membrane proteins. We uniquely detected 23 in the heavily washed insoluble fraction and 14 from the lightly washed insoluble fraction (data not shown).

At the MIPS website[24], the entire yeast genome has been analyzed for loci with predicted transmembrane (Tm) domains from 1 to 20 by applying the criteria of Klein *et al.*[34] and Goffeau *et al.*[35]. Using these criteria, 697 proteins from the *S. cerevisiae* genome have three or more predicted Tm domains, of which we identified 131 or 19% of the total (Table 2). Of these 131 proteins, 44 were identified only in the heavily washed insoluble fraction, and 33 were identified only in the lightly washed insoluble fraction. Several of these proteins have low predicted abundances based on their CAI. For example, two unique peptides were detected for the poorly characterized protein YCR017c (CAI = 0.16), which has 15 predicted Tm domains (Table 3)[24].

The peptides detected and identified from each predicted integral membrane protein rarely covered part of or all of a predicted Tm domain (Table 3). Of the 70 peptides identified from 26 proteins with 10 or more predicted Tm domains, 4 peptides partially covered predicted Tm domains (FKS1, ALG7, and YGR125w) and 4 peptides completely covered predicted Tm domains (ALG7, ITR1, PMA1, and PMA2) (Table 3). Furthermore, 43 of the 70 peptides listed in Table 3 mapped to the largest soluble domain of the respective protein. These patterns persisted with the identifications of proteins with three to nine predicted transmembrane domains.

For example, 13 unique peptides were assigned to PMA1. PMA1 is the major isoform of the H^+-transporting P-type ATPase found in the plasma membrane[36], and a three-dimensional map of a plasma membrane H^+-ATPase from *Neurospora crassa* has been reported[37]. Of the 13 unique peptides identified from PMA1 in our analysis, 10 were from the soluble-loop domain between the fourth (amino acids 326–342) and fifth (amino acids 662–678) predicted Tm domains (Fig. 3). This gap of 342 amino acids between these two predicted Tm domains is the largest domain between two Tm domains in PMA1 and is the catalytic subunit[24]. Interestingly, the peptide of amino acids 659–680, which completely covers the fifth Tm domain, was detected and identified in our analysis (Fig. 3).

An earlier comparison of the 8 Å crystal structures of both the Ca^{2+}-ATPase from sarcoplasmic reticulum[38] and the plasma membrane H^+-ATPase from *N. crassa*[37] demonstrated that the membrane domains of both of the proteins are positioned in a highly similar fashion[39]. The crystal structure of Ca^{2+}-ATPase from sarcoplasmic reticulum, a P-type ATPase, has recently been determined[40,41]. In this crystal structure, the fifth Tm domain protrudes beyond the membrane and forms a column on which the phosphorylation domain is fixed[40,41]. We detected and identified the corresponding Tm domain in PMA1 in our analysis (Tm 5 in both P-type ATPases) (Fig. 3).

Discussion

Saccharomyces cerevisiae has been the subject of a wide variety of proteomic analyses[4,5,32,42,43], but the greatest number of proteins identified previously in a single study was 279 (ref. 32). All of these studies utilized 2D-PAGE coupled to MS, which is time-consuming as a result of the nature of spot-by-spot analysis and biased against low-abundance proteins, integral membrane proteins, and proteins with extremes in pI or MW. A substitute to 2D-PAGE/MS as the method for proteomic analyses must resolve proteins as well as 2D-PAGE, allow for the rapid identification of the proteins resolved, and deal equally with proteins, regardless of their abundance, subcellular localization, or physicochemical parameters.

To achieve the resolving power of 2D-PAGE, a multidimensional chromatography method must be used. A wide variety of systems coupling multidimensional chromatography to mass spectrometry have been described[19,23,44]. Although these methods may be suitable to automation, none identified >200 proteins from any sample. Many different types of chromatography (ion exchange, reverse phase, size exclusion) may be used in tandem so long as they are largely independent and components resolved in one dimension remain resolved in the second dimension[22]. Next, a fully automated high-throughput method is needed that combines resolution and identification removing all sample-handling steps once the sample is loaded onto the system. A fully online 2D LC/MS/MS system like MudPIT fulfills both of these requirements. Once a sample is

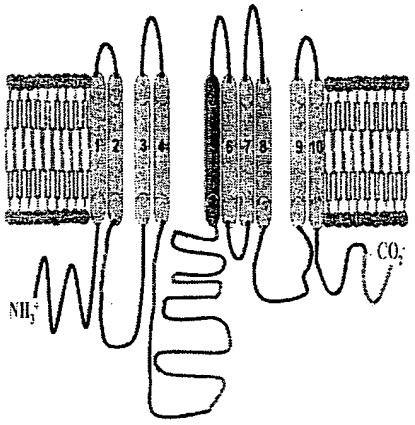

Figure 3. Peptide mapping of the integral membrane protein PMA1. A two-dimensional representation of PMA1 is displayed. Cylinders represent the predicted Tm domains as reported by MIPS (ref. 24). The protein segments between predicted Tm domains are drawn to approximate scale. Black lines and green cylinders represent segments of the protein not identified in this study. Red lines and the red cylinder represent segments of the protein identified in this study. One peptide was detected and identified between Tm domains 2 and 3, 10 peptides were detected and identified between Tm domains 4 and 5, and one peptide was detected and identified in the C terminus. We also detected and identified a peptide corresponding to Tm domain 5 in our analysis. The 320-amino acid domain between Tm domains 4 and 5 is the largest in the protein.

Table 2. Proteins identified containing three or more predicted transmembrane domains[a]

Number of predicted transmembrane domains	Number of proteins in class	Number of proteins in class identified by MudPIT	Percentage of total predicted
3	185	31	17
4	101	16	16
5	57	12	21
6	58	14	24
7	56	7	13
8	54	13	24
9	71	12	17
10	53	14	26
11	30	4	13
12	15	4	27
13	8	3	38
14	3	0	0
15	4	1	25
16	1	0	0
20	1	0	0
Totals	697	131	19

[a]The Munich Information Center for Protein Sequences website was used to obtain this information[24]. The prediction of transmembrane domains at this site is based on Klein *et al.*[34] and Goffeau *et al.*[35].

loaded onto the two-dimensional column and inserted into the system (Fig. 1), no further operator interaction is needed. The major improvement over 2D-PAGE systems is that the resolution of peptides and the generation of tandem mass spectra occur simultaneously on the same sample. That is, at any given point in time, the mass spectrometer is generating tandem mass spectra to be searched against a protein database, while the HPLC and microcapillary column are resolving and eluting peptides directly into the mass spectrometer.

Although the 1,484 proteins we identified likely do not represent a complete analysis of all the proteins present in logarithmically growing cells, our method clearly provides a large-scale and global view of the *S. cerevisiae* proteome. Our methodology not only gave access to low-abundance proteins, membrane proteins, proteins with MW in excess of 180 kDa, and proteins with pIs >10, but more importantly, it did so in a largely unbiased manner. Figure 4 illustrates this point by plotting the number of proteins identified in a particular class as a percentage of the predicted proteins. The sensitivity level across the classes of proteins listed ranged from 13% of the predicted proteins identified with pIs <4.3 and MWs <10 kDa to 43% of the predicted proteins identified with pIs >11 (Fig. 4). The method has a slight bias against proteins with a pI <4.3 and MWs <10 kDa, although proteins from both of theses classes were identified. The decreased sensitivity to these classes was likely a result of a lack of tryptic peptides in the final mixture. Generally, the smaller the protein the fewer the tryptic peptides possibly generated within the mass-to-charge ratio range of the mass spectrometer. Furthermore, proteins with pIs <4.3 have fewer lysine or arginine residues that can be targeted during the endoproteinase Lys-C/trypsin digestion. Consequently, fewer peptides are generated from those acidic proteins, decreasing their chances of being identified during a MudPIT run.

The identification of integral membrane proteins by 2D-PAGE is an intensive area of research in which progress is being made[17,18]. In the most detailed proteomic analysis of a membrane from a cell to date, Molloy *et al.* identified 21 of 26 predicted integral membrane proteins from the outer membrane of *Escherichia coli* K-12 cells[45]. We identified 131 proteins with three or more predicted integral membrane proteins (Table 2) using formic acid and CNBr as the first step in the sample treatment.

Table 3. Proteins identified with 10 or more predicted transmembrane (Tm) domains[a]

Locus	Name	No. of predicted Tm domains	No. of peptides identified	Peptide hits within Tm domains[b]	Peptide hits to largest soluble domain	CAI	MW (kDa)	Membrane localization in cell
YCR017C	–	15	2	N	1	0.16	108	–
YGR032w	GSC2	13	4	N	3	0.21	217	Plasma
YIL030c	SSM4	13	1	N	0	0.17	151	–
YJL198w	–	13	1	N	1	0.18	98	–
YDR135c	YCF1	12	3	N	2	0.15	171	Vacuolar
YKL209c	STE6	12	1	N	1	0.13	145	Plasma
YLL015w	–	12	1	N	0	0.14	177	–
YLR342w	FKS1	12	6	1 P	3	0.27	215	Plasma
YGL022w	STT3	11	2	N	2	0.21	82	ER[c]
YNL268w	LYP1	11	1	N	1	0.22	68	Plasma
YNR013c	–	11	3	1 P	1	0.19	99	Plasma
YPL058c	PDR12	11	6	N	3	0.29	171	–
YBR068c	BAP2	10	2	N	0	0.16	68	Plasma
YBR243c	ALG7	10	2	1 P, 1 C	0	0.13	50	ER
YDR342c	HXT7	10	1	N	1	0.52	63	Plasma
YDR343c	HXT6	10	2	N	1	0.52	63	Plasma
YDR345c	HXT3	10	1	N	1	0.49	63	Plasma
YDR497c	ITR1	10	1	C	0	0.19	64	Plasma
YER119c	–	10	1	P	0	0.10	49	–
YFL025c	BST1	10	1	N	0	0.13	118	ER
YGL008c	PMA1	10	13	1 C	10	0.73	100	Plasma
YGR125w	–	10	1	1 P	0	0.12	117	–
YHR094c	HXT1	10	1	N	1	0.41	63	Plasma
YLL061w	MMP1	10	1	N	0	0.13	64	–
YOR328w	PDR10	10	1	N	1	0.13	176	Plasma
YPL036w	PMA2	10	11	1 C	10	0.30	102	Plasma

[a]The Munich Information Center for Protein Sequences website was used to obtain this information. The prediction of transmembrane domains at this site is based on Klein *et al.*[34] and Goffeau *et al.*[35] [b]Abbreviations: N, none; P, partially covers a transmembrane domain; C, completely covers a transmembrane domain. [c]Endoplasmic reticulum.

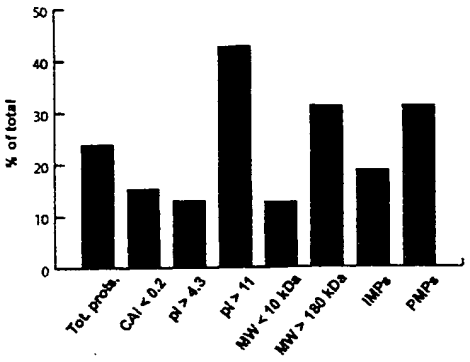

Figure 4. Sensitivity of MudPIT to a wide variety of protein classes. The percentage of proteins identified in this study from a variety of protein classes is presented. The percentages were determined by dividing the number of proteins identified in the study in each category shown by the total number of predicted proteins from each category shown. MIPS (ref. 24) and the Yeast Proteome Database[33] were used to obtain the predicted numbers of proteins from *S. cerevisiae* in each class. From left to right are the percentages identified of total proteins, proteins with a CAI <0.2, proteins with a pI < 4.3, proteins with a pI >11, proteins with a MW <10kDa, proteins with a MW >180 kDa, integral membrane proteins (IMPs) with three or more predicted transmembrane domains, and peripheral membrane proteins (PMPs).

Because formic acid is an organic acid, it partially solubilized the membrane portions of the cell in our heavily and lightly washed insoluble fractions. Then, CNBr cleaved off the soluble portions of the integral membrane proteins as large domains that were subjected to additional proteolysis. Peptides detected and identified from integral membrane proteins rarely contained any portion of a predicted transmembrane domain (Table 3). When multiple hits were obtained to a particular integral membrane protein, the peptides identified typically localized to the largest soluble loop between two predicted transmembrane domains in the protein (Table 3 and Fig. 3). In the instance of PMA1, we identified a Tm domain that may have unique functional significance (Fig. 3). On the basis of the crystal structure of another P-type ATPase (refs 40,41) and the similarities of P-type ATPases (ref. 39), Tm domain 5 in PMA1 may protrude beyond the plasma membrane and provide a column on which the catalytic domain rests. Based on the results, our method may be useful for localizing predicted integral membrane proteins to particular membranes in a cell and for providing support for predicted folding of proteins within the membrane.

Proteomics is beginning to develop the methodology needed for comprehensive high-throughput quantitative analyses of proteomes. The method described in this work is a major step toward comprehensive high-throughput methods, because not only were low-abundance proteins detected and identified, but peripheral and integral membrane proteins were also detected. MudPIT alone is not particularly quantitative. In general, the more abundant a protein, the more peptides identified from a protein. Only when emerging quantitative proteomic methods[46–49] are combined with MudPIT will true large-scale analysis of protein expression changes be possible. The combination of MudPIT with quantitative methods will allow for the integration of mRNA and protein expression levels needed to fully understand gene networks[6,7].

Experimental protocol

Materials. Standard laboratory chemicals used in this work and acid-washed glass beads were obtained from Sigma (St. Louis, MO). Sodium vanadate ($NaVO_3$), sodium fluoride (NaF), sodium pyrophosphate ($Na_4P_2O_7$), formic acid, and cyanogen bromide (CNBr) came from Aldrich (Milwaukee, WI). Poroszyme bulk immobilized trypsin was a product of Applied Biosystems (Framingham, MA). HPLC-grade acetonitrile (ACN) and HPLC-grade methanol were purchased from Fischer Scientific (Fair Lawn, NJ). Endoproteinase Lys-C was purchased from Roche Diagnostics (Indianapolis, IN). Difco Dextrose, tryptone, and yeast extract were products of BD Biosciences (Sparks, MD). Heptafluorobutyric acid (HFBA) was obtained from Pierce (Rockford, IL). Glacial acetic acid was purchased from Malinckrodt Baker Inc. (Paris, KY).

Growth and lysis of *S. cerevisiae*. Strain BJ5460 (ref. 50) was grown to mid-log phase (OD 0.6) in YPD at 30°C. To generate three fractions to analyze, two separate groups of cells were treated in the following manner. Cells were solubilized in lysis buffer (310 mM NaF, 3.45 mM $NaVO_3$, 50 mM Tris, 12 mM EDTA, 250 mM NaCl, 140 mM dibasic sodium phosphate pH 7.60) and disrupted in the presence of glass beads in a Mini-BeadBeater (BioSpec Products, Bartlesville, OK) as described[19]. After removal of the supernatants, the remaining two pellets were subjected to additional washing as follows. Each pellet was washed by adding 1× PBS (1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM dibasic potassium phosphate, pH 7.4) to the tube, vortexed for 2 min, and pelleted by centrifugation at 14,000 r.p.m. for 10 min in the Eppendorf microfuge. One pellet (to be named the lightly washed insoluble pellet) was washed once in this fashion followed by lyophilization to dryness in a Speed Vac SC 110 (Savant Instruments, Holbrook, NY). The second pellet (to be named the heavily washed insoluble pellet) was washed 3× in this way, followed by lyophilization to dryness.

Digestion of soluble fraction. After adjusting the pH to 8.5 with 1 M ammonium bicarbonate (AmBic), the protein concentration was determined by the Bradford assay. The sample was sequentially solubilized in 8 M urea, reduced by adding dithiothreitol to 1 mM, and carboxyamidomethylated in 10 mM iodoacetamide. After digestion with Endoproteinase Lys-C as described[19], the solution was diluted to 2 M urea with 100 mM AmBic, pH 8.5 followed by the addition of $CaCl_2$ to 1 mM. Finally, 3 µl of Porosyzme immobilized trypsin were added and incubated overnight at 37°C while rotating. After removal of the Porosyzme immobilized trypsin beads by centrifugation, a solid-phase extraction with SPEC-PLUS PTC18 cartridges (Ansys Diagnostics, Lake Forest, CA) was carried out on the supernatant according to the manufacturer's instructions to concentrate the complex peptide mixtures and buffer exchange the mixtures into 5% ACN, 0.5% acetic acid. Samples not immediately analyzed were stored at –80°C. After the preparation of the complex peptide mixture, amino acid analysis (Macromolecular Structure Facility, Department of Biochemistry, Michigan State University) was carried out on each sample.

Digestion of insoluble fractions. The lyophilized heavily washed and lightly washed insoluble fractions were treated separately by adding 100 µl of 90% formic acid and incubating for 5 min at room temperature. After adding 100 mg of CNBr, the samples were incubated overnight at room temperature in the dark. On the following day, the pH was adjusted to 8.5 by the addition of MilliQ H_2O and solid AmBic. Each fraction was lyophilized to ~200 µl. From this point forward, the samples were treated identically to the soluble fraction.

Multidimensional protein identification technology (MudPIT). Each sample was subjected to MudPIT analysis with modifications to the method described by Link *et al*[19]. A quaternary Hewlett-Packard 1100 series HPLC was directly coupled to a Finnigan LCQ ion trap mass spectrometer equipped with a nano-LC electrospray ionization source[51]. A fused-silica microcapillary column (100 µm i.d. × 365 µm o.d.) was pulled with a Model P-2000 laser puller (Sutter Instrument Co., Novato, CA) as described[51]. The microcolumn was first packed with 10 cm of 5 µm C_{18} reverse-phase material (XDB-C18, Hewlett-Packard) followed by 4 cm of 5 µm strong cation exchange material (Partisphere SCX; Whatman, Clifton, NJ). Approximately 420 µg of the soluble fraction, 440 µg of the lightly washed insoluble fraction, and 490 µg of the heavily washed insoluble fraction were loaded onto three separate microcolumns for the analysis of each fraction. After loading the microcapillary column, the column was placed in-line with the system (Fig. 1) as described[19]. A fully automated 15-step chromatography run was carried out on each sample. The four buffer solutions used for the chromatography were 5% ACN/0.02% HFBA (buffer A), 80% ACN/0.02% HFBA (buffer B), 250 mM ammonium acetate/5% ACN/0.02% HFBA (buffer C), and 500 mM ammonium acetate/5% ACN/0.02% HFBA (buffer D). The first step of 80 min consisted of a 70 min gradient from 0 to 80% buffer B and a 10 min hold at 80% buffer B. The next 12 steps were 110 min each with the following profile: 5 min of 100% buffer A, 2 min of *x*% buffer C, 3 min of 100% buffer A, a 10 min gradient from 0 to 10% buffer B, and a 90 min gradient from 10 to 45% buffer B. The 2 min buffer C percentages (*x*) in steps 2–13 were as follows: 10, 20, 30, 40, 50, 60, 70, 80, 90, 90, 100, and 100%. Step 14

consisted of the following profile: a 5 min 100% buffer A wash, a 20 min 100% buffer C wash, a 5 min 100% buffer A wash, a 10 min gradient from 0 to 10% buffer B, and a 90 min gradient from 10 to 45% buffer B. Step 15 was identical to step 14 except that the 20 min salt wash was with 100% buffer D.

SEQUEST analysis and AUTOQUEST output. The SEQUEST algorithm[21] was run on each of the three data sets against the yeast_orfs.fasta database from the National Center for Biotechnology Information. The AUTOQUEST software package displayed the output, listing protein loci with the number of peptides assigned to each locus. Because CNBr cleaves at methionine residues and leaves either homoserine (Hse) or Hse lactone[52], the MS/MS data resulting from the two samples treated with CNBr/formic acid had to be independently analyzed twice with SEQUEST[21]. For each run, the differential search modification was engaged and set to either –30 for Hse or –48 for Hse lactone. We used conservative criteria to determine the protein content of our samples based on those described[19]. Peptides identified by SEQUEST may have three different charge states (+1, +2, or +3), each of which results in a unique spectrum for the same peptide. Except in rare instances, an accepted SEQUEST result had to have a ΔCn score of at least 0.1 (regardless of charge state[21]). Peptides with a +1 charge state were accepted if they were fully tryptic and had a cross correlation (Xcorr) of at least 1.9. Peptides with a +2 charge state were accepted if they were fully tryptic or partially tryptic between the Xcorr ranges of 2.2 and 3.0. Partially tryptic peptides were especially relevant in the two samples where CNBr was used. Peptides with a +2 charge state with an Xcorr >3.0 were accepted regardless of their tryptic nature. Finally, +3 peptides were only accepted if they were fully or partially tryptic and had an Xcorr >3.75. We manually confirmed each SEQUEST result from every protein identified by four or fewer peptides using criteria described[19]. When five or more peptides were identified from a protein we manually confirmed that at least one of the SEQUEST results fit criteria described[19].

Note: Supplementary information can be found on the Nature Biotechnology *website in Web Extras (http://biotech.nature.com/web_extras).*

Acknowledgments

The authors thank Jimmy Eng, David Schieltz, David Tabb, and Laurence Florens for valuable discussions during the preparation of this manuscript. The authors acknowledge funding from the National Institutes of Health R33CA81665-01 and RR11823-03. M.P.W. acknowledges support from genome training grant T32HG000035-05. Saccharomyces cerevisiae strain BJ5460 was a generous gift from Steve Hahn of the Fred Hutchinson Cancer Research Center (Seattle, WA).

Received 21 August 2000; accepted 30 November 2000

1. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
2. Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K. & Matsubara, K. Expression profiles of active genes in human and mouse livers. *Gene* **174**, 151–158 (1996).
3. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533–537 (1997).
4. Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. & Garrels, J.I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
5. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
6. Hatizmanikatis, V. & Lee, K.H. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabol. Eng.* **1**, 275–281 (1999).
7. Hatzimanikatis, V., Choe, L.H. & Lee, K.H. Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.* **15**, 312–318 (1999).
8. Hanash, S.M. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. *Electrophoresis* **21**, 1202–1209 (2000).
9. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
10. Washburn, M.P. & Yates, J.R. Analysis of the microbial proteome. *Curr. Opin. Microbiol.* **3**, 292–297 (2000).
11. Langen, H. *et al.* Two-dimensional map of the proteome of *Haemophilus influenzae*. *Electrophoresis* **21**, 411–429 (2000).
12. Oh-Ishi, M., Satoh, M. & Maeda, T. Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins. *Electrophoresis* **21**, 1653–1669 (2000).
13. Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F. & Sanchez, J.C. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21**, 1104–1115 (2000).
14. Fountoulakis, M., Takacs, M.F., Berndt, P., Langen, H. & Takacs, B. Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis* **20**, 2181–2195 (1999).
15. Fountoulakis, M., Takacs, M.F. & Takacs, B. Enrichment of low-copy-number gene products by hydrophobic interaction chromatography. *J. Chromatogr. A* **833**, 157–168 (1999).
16. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional electrophoresis-based proteome analysis. *Proc. Natl. Acad. Sci. USA* **97**, 9390–9395 (2000).
17. Molloy, M.P. Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.* **280**, 1–10 (2000).
18. Santoni, V., Molloy, M. & Rabilloud, T. Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21**, 1054–70 (2000).
19. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
20. McCormack, A.L. *et al.* Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776 (1997).
21. Eng, J.K., McCormack, A.L. & Yates, J.R.I. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
22. Giddings, J.C. Concepts and comparisons in multidimensional chromatography. *J. High Res. Chromatogr.* **10**, 319–323 (1987).
23. Washburn, M.P. & Yates, J.R. Novel methods of proteome analysis: multidimensional chromatography and mass spectrometry. *Proteomics: A Current Trends Supplement*, 28–32 (2000).
24. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
25. Sharp, P.M. & Li, W.H. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
26. Peterson, C.L. & Workman, J.L. Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev.* **10**, 187–192 (2000).
27. Cairns, B.R., Kim, Y.J., Sayre, M.H., Laurent, B.C. & Kornberg, R.D. A multisubunit complex containing the SWI1/ADR6, SWI2/SNF2, SWI3, SNF5, and SNF6 gene products isolated from yeast. *Proc. Natl. Acad. Sci. USA* **91**, 1950–1954 (1994).
28. Culotta, V.C. *et al.* The copper chaperone for superoxide dismutase. *J. Biol. Chem.* **272**, 23469–23472 (1997).
29. Liu, Q. *et al.* Site-directed mutagenesis of the yeast V-ATPase A subunit. *J. Biol. Chem.* **272**, 11750–11756 (1997).
30. Lee, B.N. & Elion, E.A. The MAPKKK Ste11 regulates vegetative growth through a kinase cascade of shared signaling components. *Proc. Natl. Acad. Sci. USA* **96**, 12679–12684 (1999).
31. Sprague, G.F., Jr. Control of MAP kinase signaling specificity or how not to go HOG wild. *Genes Dev.* **12**, 2817–2820 (1998).
32. Perrot, M. *et al.* Two-dimensional gel protein database of *Saccharomyces cerevisiae* (update 1999). *Electrophoresis* **20**, 2280–2298 (1999).
33. Costanzo, M.C. *et al.* The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28**, 73–76 (2000).
34. Klein, P., Kanehisa, M. & DeLisi, C. The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815**, 468–476 (1985).
35. Goffeau, A., Nakai, K., Slonimski, P., Risler, J.L. & Slominski, P. The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett.* **325**, 112–117 (1993).
36. Ambesi, A., Miranda, M., Petrov, V.V. & Slayman, C.W. Biogenesis and function of the yeast plasma-membrane H(+)-ATPase. *J. Exp. Biol.* **203**, 155–160 (2000).
37. Auer, M., Scarborough, G.A. & Kuhlbrandt, W. Three-dimensional map of the plasma membrane H+-ATPase in the open conformation. *Nature* **392**, 840–843 (1998).
38. Zhang, P., Toyoshima, C., Yonekura, K., Green, N.M. & Stokes, D.L. Structure of the calcium pump from sarcoplasmic reticulum at 8-Å resolution. *Nature* **392**, 835–839 (1998).
39. Kuhlbrandt, W., Auer, M. & Scarborough, G.A. Structure of the P-type ATPases. *Curr. Opin. Struct. Biol.* **8**, 510–516 (1998).
40. McIntosh, D.B. Portrait of a P-type pump. *Nat. Struct. Biol.* **7**, 532–535 (2000).
41. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature* **405**, 647–655 (2000).
42. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440–14445 (1996).
43. Garrels, J.I. *et al.* Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis* **18**, 1347–1360 (1997).
44. Nilsson, C.L. & Davidsson P. New separation tools for comprehensive studies of protein expression by mass spectrometry. *Mass Spectrom. Rev.* **19**, 390–397 (2000).
45. Molloy, M.P. *et al.* Proteomic analysis of the *Escherichia coli* outer membrane. *Eur. J. Biochem.* **267**, 2871–2881 (2000).
46. Pasa-Tolic, L. *et al.* High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **121**, 7949–7950 (1999).
47. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999).
48. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
49. Münchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
50. Jones, E.W. Tackling the protease problem in *Saccharomyces cerevisiae*. *Methods Enzymol.* **194**, 428–453 (1991).
51. Gatlin, C.L., Kleemann, G.R., Hays, L.G., Link, A.J. & Yates, J.R. Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry. *Anal. Biochem.* **263**, 93–101 (1998).
52. Aitken, A., Geisow, M.J., Findlay, J.B.C., Holmes, C. & Yarwood, A. Peptide preparation and characterization. In *Protein sequencing: a practical approach* (eds Findlay, J.B.C. & Geisow, M.J.) 43–68 (IRL Press, New York, NY; 1989).